OBSERVER ERROR IN CITIZEN ORNITHOLOGY

by

Robert Gordon Farmer

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2012

© Copyright by Robert Gordon Farmer, 2012

DALHOUSIE UNIVERSITY

DEPARTMENT OF BIOLOGY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “OBSERVER ERROR IN CITIZEN ORNITHOLOGY” by Robert Gordon Farmer in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: August 2, 2012

External Examiner: ______________________________
Dr. Charles M. Francis

Research Supervisor: ______________________________
Dr. Marty L. Leonard

Examining Committee: ______________________________
Dr. Andrew G. Horn

____
Dr. Joanna Mills Flemming

Departmental Representative: ______________________________

DALHOUSIE UNIVERSITY

DATE: August 2, 2012

AUTHOR: Robert Gordon Farmer

TITLE: OBSERVER ERROR IN CITIZEN ORNITHOLOGY

DEPARTMENT OR SCHOOL: Department of Biology

DEGREE: Ph.D. CONVOCATION: October YEAR: 2012

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions. I understand that my thesis will be electronically available to the public.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in the thesis (other than brief excerpts requiring only proper acknowledgement in scholarly writing), and that all such use is clearly acknowledged.

Signature of Author

In memory of Lefty and of his untimely end.

You took one for the team, and I'm a better person because of it.

Thanks, guy.

Table of Contents

List of Tables

List of Figures

Abstract

Citizen science, which uses volunteer observers in research, is fast becoming standard practice in ecology. In this thesis, I begin with an essay reviewing the benefits and limitations of citizen science, and then measure the influence of several forms of observer error that might bias ornithological citizen science. Using an internet-based survey, I first found that observer skill level can predict the nature of false-positive detections, where self-identified experts tend to falsely detect more rare species and moderately-skilled observers tend to falsely detect more common species. I also found that overconfidence is widespread among all skill levels, and hence that observer confidence is an unreliable indication of data quality. Using existing North American databases, I then found that older observers tend to detect fewer birds than younger observers – especially if the birds' peak call frequencies exceed 6 kHz – and that published long-term population trend estimates and high-pitched ($\geq$ 6 kHz) peak bird vocalization frequencies are negatively correlated. Taken together, these data suggest that both hearing loss and other sensory changes might be negatively biasing long-term trend estimates. In the next chapter, I measured how observer experience can bias detection data. In solitary observers, I found that detections tend to increase over the first 5 years of service (e.g. learning effects), after which they decline consistently (e.g. observer senescence). Conversely, among survey groups that may be motivated to exceed a previous year's species count, I found that species richness tends to increase consistently with consecutive survey years. In this case, individual sensory deficits may be offset by group participation. Lastly, I re-evaluated the established assumption that the quality of new volunteers on North American Breeding Bird Survey routes is increasing over time. I showed that the existing measure of "quality" ignores variable lengths of observer service, and that, after accounting for this variable, "quality" is unchanging. Throughout this thesis, I also show how generalized additive mixed models can address these biases statistically. My findings offer new opportunities to improve the accuracy and relevance of citizen science, and by extension, the effectiveness of wildlife conservation and management.

List of Abbreviations and Symbols Used

Δb_j A linear regression-based, derived observer-effects covariate which corrects for relative differences in expected bird counts among BBS observers surveying the same route at different points in time (Sauer *et al.* 1994, Chapter 6)

β_2 Difference in detection probabilities on the logistic scale between observers younger than 40 and observers older than 50, for a given species (Chapter 4)

η A covariate to account for first-year-of-service effects in models of BBS data (Kendall *et al.* 1996; Link and Sauer 2002, Chapter 6)

κ A GAMM-based, derived observer-effects covariate which corrects for relative differences in expected bird counts among-BBS observers surveying the same route at different points in time. Values are derived from deviations of $\lambda_{i(j)k}$ from their route-level means (Chapter 6)

$\lambda_{i(j)k}$ A GAMM-based, random-effects, observer-effects covariate which corrects for relative differences in expected bird counts among BBS observers at particular survey routes (Chapter 6)

ω_j A hierarchical Bayesian model-based, random-effects, observer-effects covariate which corrects for relative differences in expected bird counts among BBS observers at particular survey routes (Link and Sauer 2002, Chapter 6)

BBS North American Breeding Bird Survey (Peterjohn 1994)

BCR Bird Conservation Region, a North American geographic area roughly corresponding to physiographic differences, suitable as an independent wildlife management unit (Sauer *et al.* 2003)

CBC Audubon Christmas Bird Count (Dunn *et al.* 2005)

CWS	Canadian Wildlife Service
GAM	Generalized Additive Model (Wood 2006)
GAMM	Generalized Additive Mixed Model (Wood 2006)
GLMM	Generalized Linear Mixed Model
MBBA	Maritimes Breeding Bird Atlas, available at http://www.mba-aom.ca/ accessed 5 June 2012
OBBA	Atlas of the Breeding Birds of Ontario, available at http://www.birdsontario.org/atlas/index.jsp accessed 5 June 2012
USGS	United States Geological Service

Acknowledgements

Grad school is funny business. With all its independence, it's easy to lose yourself in your own set of (too-)twisted ideas, or else to lose the day-to-day structure that keeps normal people's heads on straight. So I've got to thank all the people who kept encouraging me to talk about science and research (forcing me to see the good stuff, what isn't working, and what I don't know), to take time to get outside (fieldwork and fun), and meanwhile, to go for the long-term prospects that are going to make me happy. I am also grateful to the good people who made sure that I didn't drop the ball when I had to drop the ball.

So, thanks to Leonard Lab members, past and present, including Gabrielle Beaulieu, Greg Breed, Catherine Dale, Liz Fairhurst, Cory Matthews, Emma McIntyre, Krista Patriquin and Rob Ronconi. Thanks also to my family and to LSC affiliates Sean Anderson, Stephanie Boudreau, Kate Crosby, Neil Davies, Paul Debes, Carl Guilmette, Susan Heaslip, Aaron Heiss, Alan Hidy, Devon Johnstone, Damian Lidgard, Påûl Măttërn, Marina Milligan, Kristi O'Brien, Jackie Porter, Njal Rollinson and Christine Ward-Paige.

Thanks also to the brass: my committee, collaborators, clever superiors, and supervisors Peter Blancher, Andy Horn (who always looks his best), Ian McLaren and Joanna Mills Flemming. Enormous thanks to Marty Leonard for her patience and support for my tangents, wordiness and LaTeX.

Chapter 1

Introduction

The use of non-professional volunteers in scientific research ('citizen science') has become an accepted practice for answering broad-scale ecological questions (Bonney *et al.* 2009; Dickinson *et al.* 2010). For very little cost-per-datum, citizen science combines recreational and educational opportunities with data collection over unprecedented geographic and time-scales, and, when married with modern internet systems, can rapidly characterize changing environmental systems to aid in periodic assessments and adaptive management (Wiersma 2010). For some applications, for instance monitoring populations of many bird species (Sauer and Link 2011), citizen science projects (*e.g.* the North American Breeding Bird Survey; Peterjohn 1994) provide the principal or even the only dataset that is useful to managers. With ecosystems undergoing rapid, contemporary changes, often due to manageable anthropogenic influences (*e.g.* Child *et al.* 2009; Barnosky *et al.* 2011), it is thus critical that citizen science, including the accompanying data analysis, be both accurate and comprehensive in order to avoid inappropriate management decisions (*e.g.* Thomas and Martin 1996; McKelvey *et al.* 2008; Lindenmayer and Likens 2010).

However, with a large and often less-supervised set of observers than conventional studies, citizen science projects are subject to a poorly-understood set of observer errors (Dickinson *et al.* 2010). Some notable, potential sources of observer error in citizen science projects include flexible survey designs which promote recreational objectives, potentially at the expense of methodological rigour (Dunn *et al.* 2005); the use of a large pool of volunteer participants, often with inconsistent skill levels that are not accounted for (Fitzpatrick *et al.* 2009); and the long-term participation of individuals, who might not collect data consistently from year to year (Kendall *et al.* 1996). The effect of these errors on data quality, and how they might be mitigated, requires further study. Along these lines, because citizen science often makes use of a wide and changing variety of statistical methods (*e.g.* Fewster *et al.*

2000; Kéry and Royle 2010; Sauer and Link 2011), there are regular opportunities to improve data collection and analysis methods to better account for such errors (*e.g.* Kendall *et al.* 1996; Miller *et al.* 2011; Sauer and Link 2011).

The purpose of this thesis is to recognize, measure, and develop mitigation strategies for observer errors in citizen science. Here, I analyzed data from citizen science datasets including the North American Breeding Bird Survey ('BBS' Peterjohn 1994), the Audubon Christmas Bird Count ('CBC' Dunn *et al.* 2005), and the Atlas of the Breeding Birds of Ontario ('OBBA' Bird Studies Canada *et al.* 2008). I chose to work with ornithological datasets because at present, avian science has a multi-century history of citizen participation (Greenwood 2007) and the data available are some of the most extensive among any form of citizen science (Dickinson *et al.* 2010). My focus on observer error in ornithological datasets is thus both convenient and highly-relevant to current ecological management. While my results and conclusions are most relevant to bird surveys, some concepts, for instance the effects of skill level and of observer senescence on detection probability, easily apply to other fields which involve multiple observers conducting field counts, including plant ecology (Fitzpatrick *et al.* 2009) and herpetology (de Solla *et al.* 2005).

In Chapter 2, I review the definition and some examples of citizen science, and present what I believe to be the most effective designs and analysis principles for citizen science and similar monitoring projects. In doing so, I consider the problems of imperfect detection, historical and recreational constraints to survey designs, and variations (and perceived variations) in observer quality. Each of these sources of error can affect the quality of collected data and any biological conclusions arising. I then discuss how citizen science surveys can be designed to efficiently achieve specific research goals, but further argue for the value of long-term monitoring in some cases – which incorporates deliberate inefficiency – in order to achieve both broader public participation in ecological science, and to give us future opportunities to make long-term or serendipitous comparisons. This essay provides a useful context for the rest of my research.

In Chapter 3, I focus on signal detection by observers identifying birds by ear. I conduct an experiment designed to determine how observer skill, competitive incentives to identify rare species, and observer confidence affect the likelihood of making correct and incorrect detections of various bird species' songs. I also measured the reliability of a birdwatcher's subjective "confidence" when recording detections. This experiment addresses potential weaknesses of citizen science projects, including involving observers of different skills, promoting competition among participants to detect rare species, and relying upon arbitrary measures of observer confidence in order to estimate correct- and incorrect detection probabilities. In this experiment, I found important skill-dependent differences in the detection of rare and common species, as well as a widespread overconfidence among observers of all skill levels. These results are important contributions to the design and analysis of auditory surveys.

In Chapter 4, I consider the consequences of multi-year service histories among volunteer observers, in particular, the potential for changes within observers over time. This concept is a recurring theme for the remaining chapters. In this particular chapter, I evaluate the colloquially-acknowledged, but nonetheless poorly-researched notion that older observers are more prone to hearing loss, and hence detect fewer birds than younger observers. Undetected hearing losses (or other sensory declines) might indicate an apparent decline in birds over time that does not correspond to real biological change. It is important to determine the extent of this potential source of error in modern survey data, since it might erroneously mobilize conservation resources for bird populations presumed to be declining or at risk of extinction, but which in fact are healthy (sensu Thomas and Martin 1996).

Using two independent datasets and two novel analysis methods, I found evidence that older observers detect fewer birds than younger observers, and I also show how some detection declines occur with increasing frequency in species with increasingly higher-pitched songs, which is suggestive of hearing loss effects. I then illustrate how existing population trend estimates might be correlated with the corresponding species' call frequencies, suggesting that hearing losses may have already affected some estimates of avian population trajectories. This research draws attention to a source of observer error which is not routinely addressed in the published literature, but which appears to be quite influential. It also uses a relatively new statistical technique, generalized additive mixed modeling (GAMMs; Wood 2006), and demonstrates how it can account for longitudinal changes in observer ability such as the sensory declines in question.

In Chapter 5, I consider how years-of-service on a given survey route, like observer age, might be a source of error. Although observer age and years-of-service are closely-correlated, some observers might serve on multiple survey routes sequentially – for instance in the North American Breeding Bird Survey ('BBS'; Peterjohn 1994) – and so its relationship to detection ability is not necessarily identical. Furthermore, previous research has shown 'start-up' effects occurring in the first years of observer service (Kendall *et al.* 1996; Jiguet 2009; Eglington *et al.* 2010) which warrant future investigation using GAMM techniques here. Lastly, years-of-service are much more-easily calculated retrospectively than observer age, and so are a more useful focus for statistical corrections in the future. Hence, alongside observer age effects, years-of-service effects are an important, additional source of error worthy of investigation. Here, I observe what are apparently both start-up learning effects, as well as the later-term sensory declines already described in Chapter 4. I reveal a more complex picture of years-of-service effects than what has been conventionally recognized, including a more prolonged initial learning curve (*i.e.* 5 years vs. 1 year), and a steep, opposing pattern of declining expected counts in later years. I also consider data from the Audubon Christmas Bird Count ('CBC'; Dunn *et al.* 2005) and highlight a source of observer error that has not been previously recognized which may be unique to surveys conducted by multiple (*vs.* single) observers – namely an inflation of species richness with increasing continuous years of service. By enhancing our understanding of existing correction factors and showing processes that have not previously been measured in the CBC, this chapter provides useful ways to improve the accuracy of model inferences.

In my final chapter (Chapter 6), I synthesize insights from Chapters 4 and 5 to provide a new and robust interpretation of a known relationship between an observer's year-of-first-service ('start year') and his or her expected counts. Previous research (Sauer *et al.* 1994) has noted that coefficients accounting for differences in expected counts among BBS observers tended to increase among observers with more-recent start years. Some authors have interpreted this pattern to mean that there has been a systematic increase in BBS birdwatcher quality over time (Sauer *et al.* 1994; Dunn *et al.* 2005). I show how this pattern is more likely driven by these coefficients' underlying relationship to observer years-of-service, where longer-serving observers (with less-recent start years) tend to be older, and so have lower expected counts for physiological reasons. Hence, the positive relationship between expected counts and an observer's start year does not necessarily reflect a systematic increase in among-observer 'quality' in recent years, but rather a tendency for within-observer detection ability to decrease over time. This research clarifies the origin of an otherwise-unusual pattern, and further underscores the importance of accounting for long-term years-of-service effects when making inferences from long-term, multi-observer surveys.

This thesis reveals important new patterns of observer error affecting several citizen-science surveys, including variations due to skill (Chapter 3), age (Chapter 4), and years of service (Chapters 5 and 6). Using case studies in many cases, it also shows how these errors might bias inferences about real population trajectories, and so prevent appropriate wildlife management from taking place. Working with some relatively novel statistical techniques, including GAMMs, this research also provides useful approaches to accounting for such errors, and has relevance to many other ecological subdisciplines which also must manage the influences of observer error. Collectively, this research thus provides important opportunities to realize the full potential of citizen science.

1.1 Publications Arising from the Thesis and Copyrights

Chapters 3, 4, 5 and 6 are written as manuscripts, including coauthors as needed, and using "we" pronouns throughout. Details of the author contributions are described in the "Student Contribution to Manuscripts in Thesis" form accompanying this thesis; however I was the principal designer, data collector, data analyst and writer in each case.

Chapter 3 has already been published as:

> Farmer, R. G.; Leonard, M. L. & Horn, A. G. (2012). Observer effects and avian call count survey quality: rare-species biases and overconfidence. Auk. **129**(1):76-86. doi:10.1525/auk.2012.11129

The publication agreement between the Auk (journal) and me allows me to republish the article “as part of any book or anthology, of which [I am] the author or editor, unless the anthology is drawn primarily from The Auk”, so long as I acknowledge the original publication. Since I have just done so, I do not need to seek further permissions to include it in this thesis.

Chapters 4, 5 and 6 will be submitted to other journals, but at present, have not yet been published.

Chapter 2

Optimizing Citizen Science and Avian Monitoring

2.1 Overview

Citizen science, which can be generally defined as the involvement of volunteers in research (Dickinson *et al.* 2010), has become an increasingly important component of modern ecological research and ecosystem management, especially for the study and conservation of birds. For example, more than 70% of the ornithological monitoring effort in Great Britain is conducted by volunteers numbering in the thousands (Battersby and Greenwood 2004; Bell *et al.* 2008). North American volunteer-based surveys such as the North American Breeding Bird Survey ('BBS'; Peterjohn 1994) and the Audubon Christmas Bird Count ('CBC'; Dunn *et al.* 2005) enlist similarly vast and growing numbers of volunteers annually (*e.g.* 60,000 to 80,000 volunteers on the CBC; Cohn 2008). These surveys, as well as similar atlassing projects (reviewed in Donald and Fuller 1998; Gibbons *et al.* 2007) benefit enormously from such massive, often continental-scale volunteer efforts, along with in-kind support (*e.g.* transportation costs), survey fees and society membership dues (Battersby and Greenwood 2004; Schmeller *et al.* 2009). Such benefits are reflected in the initiation of a growing number of citizen science projects worldwide (reviewed in Greenwood 2007; Silvertown 2009), with designs that continue to evolve alongside advances in communications technologies (*e.g.* internet-based data entry; Sullivan *et al.* 2009) and in response to developments in methods research (*e.g.* Kéry and Schmid 2004).

Numerous, recent reviews of the history and use of citizen science already exist (Donald and Fuller 1998; Greenwood 2007; Dickinson *et al.* 2010). To complement this research, here I consider how such monitoring projects can best collect large amounts of high-quality information. In this light, I consider the strengths and weaknesses of (1) controlling for error using simple versus complex designs and models, and (2) using 'amateurs' as data collectors. I then consider (3) how to maximize a survey's statistical power and efficiency, and whether the inherent inefficiencies of long-term 'surveillance' monitoring project designs lacking specific hypotheses and conceptual models (*e.g.* many citizen science projects) make them inferior to more targeted study designs. I argue that (i) complexity in both survey design and modeling is generally important for survey data quality, but that (ii) exclusive survey designs which depend upon observer judgment and expertise are problematic. More broadly, I also argue that (iii) surveillance monitoring surveys designed without specific, mechanistic hypotheses in mind can contribute high-quality scientific knowledge under certain conditions. Below, I address each of these ideas in turn.

2.2 Controlling for Error Using Designs and Models

To appreciate how survey data quality might be optimized, one must first recognize important sources of error and how they can be minimized or removed. In general, all long-term ecological surveys consist of observational data collected sequentially at known times and locations. Barker and Sauer (1992; cited in Thomas [1996]) describe how such raw data are composed of four additive processes, namely trend (prevailing tendency), interventions (*e.g.* weather events), autocorrelation and sampling error. The goal of a survey designer is to minimize the amount of the latter two nuisance effects (or the latter three, in the case of interventions [*e.g.* weather] having only transient impacts). The goal of a modeler is to remove whatever error remains. Ultimately, the aim is to reduce the amount of nuisance variability to below that of the real trend's variability, so that biological patterns can be recognized (sensu Johnson 2008). In most cases, the goal is not (or should not be) to eliminate all bias – at least because overfitted models tend to be less precise (*e.g.* James *et al.* 1996; Link and Sauer 1997*a*) – but rather to reduce that bias to a tolerable level for the given research question at hand (Elphick 2008). However, the most effective designs and modeling approaches in this regard tend to be complex, and this complexity can limit their widespread use. Below, I discuss some features of effective survey and modeling approaches, as well as some practical issues concerning their adoption, including complexity.

2.2.1 Design and Model Complexity: The Importance of Accounting for Detectability

All else being equal, design-based approaches to minimizing error are always preferred to model-based approaches in that the error is essentially prevented rather than corrected for (Bart *et al.* 2003; Johnson 2008). This contributes to increased inferential precision. In practise, however, useful inferences usually result from a combination of these methods (*e.g.* Sauer *et al.* 2004). In increasing order of statistical rigour, designs for sampling avian data include (1) anecdotal observations (one component of the 'eBird' survey; Sullivan *et al.* 2009), (2) fixed-area surveys by one or several observers incorporating some measure of effort (*e.g.* the CBC, some atlasses), (3) unlimited-radius point counts by a single observer for a fixed effort (*e.g.* the BBS), and (4) approaches that account for imperfect species detectability, including distance sampling (Newson *et al.* 2008), double-observer methods (Nichols *et al.* 2000) and time-to-detection methods ('removal models'; Farnsworth *et al.* 2002; Riddle *et al.* 2010). Repeated sampling of a site by a single observer is another powerful option that can account for false-positive detections along with missed detections if the data and models are appropriate (*e.g.* Royle and Link 2006; McClintock *et al.* 2010*b*; Miller *et al.* 2011), for instance if the frequency of false positives is low (Campbell and Francis 2011), and there is a known measure of uncertainty associated with each data point (Miller *et al.* 2011). Finer details of most of these designs are reviewed in Elphick (2008), Nichols *et al.* (2009) and Miller *et al.* (2011).

Controlling for detection probability is a key element of more-sophisticated designs. In general, the terms 'presence/absence' and 'population counts' as applied colloquially to survey data are misnomers; count records in fact represent a detection/nondetection process (usually) conditional on species presence (MacKenzie *et al.* 2005; and see Diefenbach *et al.* 2007). Hence, in the absence of controls for detection error (*e.g.* most designs of styles 1–3, above), count data are an accurate index of species populations only if the detection probability is roughly constant between sampling events (Johnson 2008). Past research has shown that detectability can vary with species (Diefenbach *et al.* 2003), time of year (Dennis *et al.* 2006), study habitat (Hanowski and Niemi 1995; Gu and Swihart 2004), survey method (Dunn 1995; Riffell and Riffell 2002), and effort (Link and Sauer 1999; Ferrer *et al.* 2006; Bonardi *et al.* 2011), among other factors, and so the assumptions of these simpler survey designs are unlikely to be met in many field situations. In most cases, it is therefore best to account for detection error using a more complex design (*i.e.* of style 4).

Model-based covariates can provide additional corrections which cannot (or did not) occur by design. These covariates can take the form of fixed- or random-effects coefficients, and can include fixed-effects covariates for background noise (Pacifici *et al.* 2008; Griffith *et al.* 2010), effort (Link and Sauer 1999), and years of service (*e.g.* Kendall *et al.* 1996, and see Chapter 5); and normally distributed random-effects covariates for individual observers (*e.g.* Link and Sauer 2002). An even more-sophisticated approach known as hierarchical occupancy modeling divides Y (*i.e.* the observed data) into separate functions for occupancy and detection (conditional on occupancy), each of which can contain relevant covariates (Royle *et al.* 2005). Using this flexible format, one can measure and account for covariables affecting detection, such as observer age (Chapter 4), or for covariables affecting occupancy, such as predictors of local extinction within a broader metapopulation (Royle and Kéry 2007). This approach can aid management planning, for instance by identifying risk factors for nondetection (*e.g.* older observers; Chapter 4), or by identifying source and sink populations in a broader landscape (*e.g.* at landscape edges; Royle and Kéry 2007).

While it is important to not overparameterize a model (*i.e.* to include too many covariates; Burnham and Anderson 2002), an interesting argument between Bart *et al.* (2003) and Sauer *et al.* (2004) highlights the importance of there being some routine level of model complexity. In this case, Bart *et al.* (2003) suggested that it might be possible to both improve and simplify the analysis of BBS data using a modeling approach that relies more heavily on design controls and less on corrective covariates. As a result of its using fewer parameters, this approach increased model precision. However, in a subsequent rebuttal to criticisms provided by Sauer *et al.* (2004), Bart *et al.* (2004*b*) qualified their earlier claim by recommending that standard (*i.e.* more-complex) models be routinely compared against these simpler models to first evaluate whether important biases are missed. In other words, in their response, Bart *et al.* (2004*b*) largely conceded that in spite of the loss of precision, more complex approaches have greater value under typical conditions.

Using a combination of design-based and appropriate modeling approaches, data quality can be maximized by recognizing important variation in detectability either implicitly (as detectability estimates made possible by appropriate survey designs), or explicitly (as model-based covariates), provided that the designs and models used are of sufficient complexity. Unfortunately, because the best statistical approaches tend to be more difficult to use, require large sample sizes, and require supplementary data (*e.g.* distance sampling detection curves; Johnson 2008), many are practically unfeasible. Lower-quality biological estimates resulting from limited statistical expertise among project staff (*e.g.* 'black-box' use of statistical tools; Johnson 2008) or small sample sizes can often be overcome with additional funding and sampling; however 'legacy' issues resulting from historical design limitations (*e.g.* Dunn *et al.* 2005; Freeman *et al.* 2007), can be more intractable.

2.2.2 The Restrictive Legacy of Some Designs

'Legacy' limits on the quality of future inferences can arise when historical data collected without a more complex (future) modeling method in mind are not suitable for the modern approach. For example, because our knowledge of the importance of imperfect detection probability has only been popularized for the past few decades (*e.g.* Kéry and Schmid 2004), many older surveys (*e.g.* the BBS [1966 to present] and the British Common Birds Census [ca. 1960–2000; Freeman *et al.* 2007]) did not collect sample replicates and supplementary data which might be used to account for such an error, and so their data cannot be modeled with imperfect detection in mind. Growing concerns about other factors such as false-positive detections (*e.g.* Royle and Link 2006; McClintock *et al.* 2010*b*; Miller *et al.* 2011, and see Chapter 4) suggest that even recently-implemented survey designs, for instance the French Breeding Bird Survey (2001 to present; Jiguet 2009) and the Swiss Survey of Common Breeding Birds (1999 to present; Royle *et al.* 2007), could be further improved, for instance by collecting measures of observer certainty (Miller *et al.* 2011). Where such design limits are present, how can survey data quality nonetheless be maximized?

First, some improved statistical approaches are still suitable for use on old datasets. For instance, advances in desktop computing power have allowed hierarchical Bayes designs to replace among-strata weighting schemes in models of (older) BBS data

(Sauer and Link 2011). These techniques can provide some improved – if not ideal – inferences from legacy information, making the best of a bad situation. Nonlinear modeling techniques (*e.g.* James *et al.* 1996), and especially generalized additive models ('GAMs'; Fewster *et al.* 2000; Clarke *et al.* 2003; Flemming *et al.* 2010, and see discussion in Kery and Royle [2010]) and their more-powerful complement, generalized additive mixed models ('GAMMs'; Wood 2006; Zuur *et al.* 2009, and see Chapters 4 and 5), are other developing statistical options that show tremendous promise for improving our understanding of long-term biological patterns from both 'legacy' and current survey data.

Compared to the current fixed-parameter approaches, additive models can more naturally illustrate simultaneous population and covariate 'trajectories' (*i.e.* nonlinear patterns of change; Link and Sauer 1997*b*) alongside 'trends' (*i.e.* averaged log-linear measures of change between two arbitrary endpoints; Link and Sauer 1997*b*). The ability to visualize such trajectories is important for appreciating whether a variable's significantly increasing or declining patterns are sensitive to the choice of baseline (which itself could have been exceptional; *e.g.* a harsh winter; Thomas 1996) and timeframe (Dunn 2002; Magurran *et al.* 2010), and these approaches are generally applicable to both older and newer datasets.

However, in what might be a limitation of having too many complex statistical tools at present, there is sometimes no clear consensus on what is the 'appropriate' modeling technique for the same legacy dataset. In the case of the BBS, Thomas and Martin (1996) showed how the separate, similarly-complex approaches taken by the Canadian Wildlife Service and the United States National Biological Service (now administered in title by the United States Geological Service) can lead to different biological conclusions, notably the value and significance of population trend estimates. Choosing and justifying a particular 'optimal' modeling method is thus a very important exercise; sensitivity analyses with several different methods can help to qualify their robustness (*e.g.* Sauer and Link 2011).

Unfortunately, once the optimal methods are decided-upon, making the corresponding improvements to the designs and analysis techniques of legacy surveys can still be quite challenging, since major changes can limit the future use of the older data by essentially introducing a discontinuity into the overall dataset. Illustrating one successful approach to this problem, in the United Kingdom, the older Common Birds Census, which began in the early 1960s, suffered from geographic biases, where effort was more concentrated in southern England and census sites were distributed nonrandomly over space (Freeman *et al.* 2007; Magurran *et al.* 2010). This led to unreliable overall and regional trend estimates, and motivated the eventual abandonment of this survey in favour of a new, better-designed survey (the UK Breeding Bird Survey) in the 1990s. To accommodate this major change from one survey design to another, both surveys were conducted in parallel for seven years. Afterwards, trend estimates for each approach showed good consistency, which, fortunately, recognized that decades of the earlier, albeit problematic, work still had much value (Freeman *et al.* 2007). With the support of funders, managers and volunteers, the transition to an improved monitoring scheme in this case proceeded smoothly, legacy issues were removed, and the survey dataset could maintain its ecological relevance.

Depending on a survey's history and user base, some design improvements can be slow to implement. The Audubon Christmas Bird Count ('CBC') provides an interesting case study of difficulties associated with changing an originally non-scientific survey into a more-scientific one. Like the UK Common Birds Census, managers have been aware for many years of problems with its data collection approaches (Francis *et al.* 2004; Dunn *et al.* 2005), but many critical changes, for instance stratifying observed counts by different methods used (such as walking or feeder-watching), are still pending. Major changes to this legacy design may come into conflict with the strong social component of CBC participation, and could lead to a significant and undesirable drop in participation (Dunn *et al.* 2005). The CBC was originally formed in 1900 as a recreational alternative to hunting, not a formal biological census, and many of its participants value the annual traditions associated with it (*e.g.* Bonta 2010). While some problems can presently be accounted for using model-based approaches and no changes to the field protocols (*e.g.* modeling nonlinear effects of effort; Link and Sauer 1999; Lepage and Francis 2002), others, especially the CBC's limited documentation of different data collection methods used, will continue to constrain the usefulness of survey data for scientific purposes (Francis *et al.* 2004, Figure 2.1). The present challenge is to implement the outstanding, necessary changes to the survey without undermining its traditional and recreational values.

In sum, increasing design and model complexity is generally important for recognizing the real complexity of the detection process, including the effects of imperfect detection and among-observer differences, both of which are common to citizen science projects (*e.g.* Sauer *et al.* 1994; Kéry and Schmid 2004). Design- and model-based approaches play equally valuable roles in removing these sources of error from the underlying biological trend, but sometimes, ideal modeling methods may not be feasible because of the methodological and cultural legacies of older survey designs. While statistical and technological developments have provided new options for improving the quality of inferences from older datasets, the best-quality information often arises from patterns observed using multiple analysis methods.

Now that I have established some advantages of more complex survey protocols, I next address whether surveyors need to be similarly 'advanced' in order to contribute meaningful monitoring data.

2.3 The "Amateur": an Asset, or a Liability?

While citizen science projects tend to involve enviably large sample sizes, there is a lingering concern about the quality of the data they can produce. This concern largely arises from the perception that the skill level of observers is less consistent – or else lower – among citizen science projects than among smaller, professionally-staffed studies (Greenwood 2007). Even if automated quality-control mechanisms are effective, for instance in identifying bird records that are unusual for a particular time and/or location (Sullivan *et al.* 2009), for certain protocols, subtler errors such as missed detections can still occur among novice participants that will not be noted (*e.g.* McLaren and Cadman 1999). To successfully involve large numbers of participants also requires less-controlled protocols that accommodate the more erratic availability of volunteers (Magurran *et al.* 2010), and which can consequently limit modeling options.

Are the more-numerous data produced by 'amateur volunteers' therefore of a lower quality than what is collected by professional scientists (see Ellis and Waterton 2005; Greenwood 2007; Cohn 2008)? Similarly, is citizen science a lesser cousin to 'professional' science (*e.g.* Cohn 2008)? Here, I argue how 'amateur' status can be irrelevant to data quality, and how skill level should also be largely unimportant to a well-designed protocol, with few exceptions. On the other hand, I briefly highlight an unfortunate tradeoff between volunteer surveyor participation and data quality.

First, I address the concept of the 'amateur' in scientific research as it relates to ability and data quality. In the context of "citizen science", the word 'citizen' is synonymous with 'volunteer amateur' (sensu Bell *et al.* 2008; Cohn 2008). Hence, citizen science data are collected largely by amateurs. The colloquial distinction between an 'amateur' and a 'professional' is one of skill; that is, amateurs are less-skilled than their paid, 'professional' colleagues (sensu Greenwood 2007). If this distinction is correct, because skill is an important determinant of the error rates in species identification data (*e.g.* McLaren and Cadman 1999; Fitzpatrick *et al.* 2009, and see Chapter 3), the implication is that citizen science data derived from species identifications must by definition be of a lower quality than what are professionally collected (*e.g.* Fitzpatrick *et al.* 2009).

However, if an 'amateur' designation is technically a function of professional affiliation, it can be largely independent of skill (*e.g.* Lotz and Allen 2007). This is especially likely for many types of field ecology – in particular field ornithology – that are easily practised recreationally by the general public (Mayfield 1979; Silvertown 2009). In many cases, local 'amateur' experts are often as equally-skilled as professionals, and may also be more knowledgeable about their particular study sites (Greenwood 2007). Similarly, many 'professional' scientists may also conduct 'amateur' surveys in their spare time (Ziolkowski Jr. and Pardieck 2006). In this way, 'amateur' 'citizen science' data can be equivalent or superior in quality to 'professional' data. Thus, the risk of lower data quality need only apply to cases when non-experts (and not non-*professionals*) are predominantly involved in citizen science projects.

Can non-expert amateurs also contribute high-quality data to scientific monitoring projects? Under the best project designs, the answer is yes. This is because much of the difficulty of doing 'science' per se should be associated with its design, analysis and interpretation, and not its data collection. In contrast, study methods must be understandable and repeatable by all outside researchers in order for important findings to be reproduced. In this regard, non-experts should be able to make valuable data-collection contributions to broad-scale monitoring projects if the

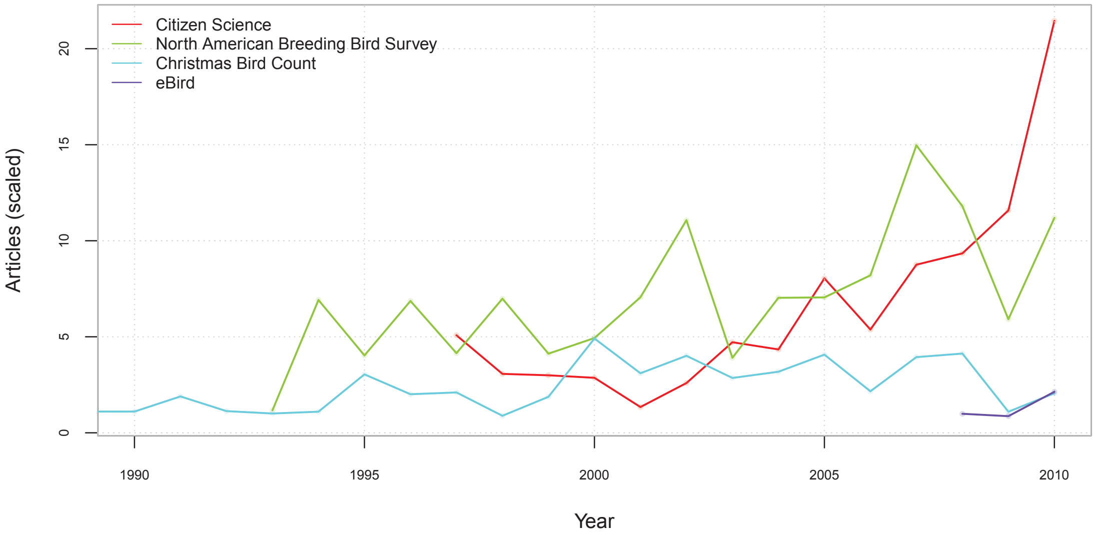

Figure 2.1. Number of records, 1990-2010 returned by the ISI Web of Science database for the corresponding blocked search terms, arranged by publication year, and scaled by the number of articles containing the keyword "ecology" for each year ($\frac{x}{x_{ecology}} \cdot 5000$). The North American Breeding Bird Survey ('BBS') is a more tightly-controlled survey design, whereas the Christmas Bird Count ('CBC') and eBird have fewer effort controls.

project methods are properly explained and specific, and if no special equipment is needed (Mayfield 1979; Greenwood 2007). The recent explosion of citizen science data accepted in peer-reviewed scientific literature undoubtedly involves many non-expert participants (Figure 2.1), and so attests to this idea holding true (*e.g.* Cohn 2008; Silvertown 2009; Ryder *et al.* 2010).

However, in some exceptional cases – frequently in 'omnibus' multi-species surveys – study methods can demand a high degree of expertise from data collectors. For example, rather than requiring participants to observe common bird feeder visitors (*e.g.* Project FeederWatch; Lepage and Francis 2002), report anecdotal observations of common birds (*e.g.* eBird; Sullivan *et al.* 2009), or count chicks in nests (*e.g.* Neighborhood NestWatch; Evans *et al.* 2005) – each of which are tasks requiring little expertise – surveyors of the BBS must in most cases be able to exhaustively distinguish from among more than 50 bird species that are likely to be heard or seen on a given survey route, a skill which tends to develop only with years of practice.

Unfortunately, even with the greater expertise of their volunteers, omnibus surveys are still problematic in that the subjective judgement of their expert observers plays a major role in determining the data (*i.e.* particular species identifications and counts made by ear); this can make results difficult to reproduce precisely (*e.g.* Robbins and Stallcup 1981; Hull *et al.* 2010), which runs contrary to the general notion that scientific results should be independently verifiable. Other known errors with omnibus survey data, also associated with differences among and within observers (Link and Sauer 1998, and see Chapters 3 to 6), add to the problems associated with this subjectivity. At present, however, there are few alternatives for collecting such specialized information. Accordingly, I consider omnibus surveying to be a flawed, if inevitable methodological approach at present.

Fortunately, omnibus survey designs that rely too heavily on subjective expertise might soon be made obsolete by technological improvements. Recent experiments using off-site expert- or computer-assisted determination of species density (Dawson and Efford 2009) and species composition (Haselmayer and Quinn 2000; Rempel *et al.* 2005; Campbell and Francis 2011) from audio recordings promise a greater amount of objectivity and increased, useful participation by non-experts (*i.e.* as recorders) in future omnibus monitoring, and by extension, in citizen science as a whole.

Nonetheless, while all 'amateurs' should ideally be able to provide high-quality scientific data, experimental conditions in citizen science (*e.g.* time of day, observation period, number of replications) tend to be somewhat less-controlled in order to accommodate the different schedules and spatial availability of volunteers compared to paid staff. This means that in spite of the equivalent data-collection potential among volunteer amateurs, the more participants a survey accommodates, the less-useful its data can become. For instance, while data are easily collected anecdotally by many volunteers during their spare time, this information is not well-controlled for effort effects and spatial biases, and so cannot be used for most statistical analyses (Dunn *et al.* 1996, 2001; McKelvey *et al.* 2008). In contrast, the longer blocks of dedicated survey time that are usually required by a monitoring project's more restrictive – but more statistically defensible – design cannot be accommodated by as many otherwise willing participants.

Intermediate time commitments such as annual surveys are evidently manageable by many volunteers, but as discussed in the first section of this review, there is a major statistical advantage to having replicated data collected by the same observer within the same year (*e.g.* Royle and Dorazio 2009). Unfortunately, the corresponding increase in required time commitments for replicated sample designs will tend to thin the pool of willing volunteers (sensu Dunn *et al.* 2005). Whether such a reduction in available participants seriously affects the viability of a project is therefore an important, design consideration, and a major limitation of citizen science data. The advantages of broader-scale data collection may or may not outweigh the disadvantages of simpler data collection protocols (*e.g.* Peterjohn 2001; Schmeller *et al.* 2009), but there is no one 'best' approach: this ideal level of volunteer participation (and protocol rigour) is specific to a given project (*e.g.* Snall *et al.* 2010).

In general, well-designed citizen science projects are insensitive to the amateur or professional status of their participants, and with a few exceptions (*e.g.* omnibus surveys) should also be insensitive to their skill levels. The only important reasons to consider citizen science to be generally inferior to professional science are if its protocols are simplified to the point of compromising the desired level of detail (*e.g.* using the BBS for a species-habitat association study, for which its lower resolution is unsuited; Peterjohn 2001), or if it cannot recruit sufficient volunteers to survey a particular (often remote) area (Francis *et al.* 2009).

2.4 Designing Surveys for Power and Efficiency

So far, I have argued that the quality of citizen science data benefits from design and model complexity, that it is generally insensitive to whether its participants are amateurs, and that it can suffer from oversimplified protocols that enhance volunteer participation. I now take a final and broader perspective and consider the statistical power and efficiency of surveys, including how they might be maximized by design. I also consider a special case where efficiency might be deliberately ignored as a design priority, asking whether monitoring projects should always include targeted research interests with mechanistic hypotheses in mind, or whether some less-focused 'surveillance' project designs of lower efficiency might also be justified.

2.4.1 Survey Design, Statistical Power and Efficiency

Two important endpoints of survey quality are its statistical power and its efficiency. First, in population-monitoring surveys, statistical power is the ability of a design to detect a certain amount of significant change a certain percentage of the time. A high-powered survey, for example, might detect a 25% change in a population over 20 years, 80% of the time (sensu Bart *et al.* 2004*a*), whereas a more modestly-powered survey might only detect such a change 60% of the time. Power to detect change increases with monitoring intensity (Purcell *et al.* 2005; Thogmartin *et al.* 2007), which means that multi-species ('omnibus') 'monitoring'-type studies, that survey tens of species at a time using a generic protocol, can be less powerful per species than more specialized programmes targeting specific species or specific species groups (*e.g.* Sauer *et al.* 2008, 2010).

For a given survey design, how can statistical power be maximized? In general, the simplest approach is to increase sampling intensity as funding and logistics allow (all else being equal; *e.g.* Bart *et al.* 2004*a*; Francis *et al.* 2005). However, for existing surveys, this approach can be expensive and thus inefficient compared to design changes (*e.g.* field protocol changes; Sauer *et al.* 2005). Field *et al.* (2007) recommend reducing the desired value of alpha for the statistical tests that are used; in other words, to require a lower burden of proof for demonstrating 'significant' change. This approach, which can be quite useful for conservation-based monitoring, simultaneously decreases the number of missed, real changes (*i.e.* increases the power), and can lead to cost savings if the cost of reacting to false alarms is low compared to the cost of failing to respond to real population changes (Field *et al.* 2004).

Survey efficiency, which is analogous to the relative amount of useful data collected for a given effort, is another important factor to be maximized once a minimum level of statistical power has been met (*e.g.* Pierce and Gutzwiller 2004). One important trade-off in this case exists among different designs for omnibus monitoring surveys, which should ideally detect changes in indicator variables for as many species as possible, for as little effort as possible. Here, the most efficient monitoring approach is different for rare species than for common species. Specifically, because rare species tend to have low abundances at any given survey site, year-to-year fluctuations in observed counts as a proportion of original count values tend to be quite high and non-significant. Consequently, for broad-scale studies of rare species, detection/nondetection surveys tend to be more efficient than count surveys (MacKenzie 2005; Joseph *et al.* 2006; Pollock 2006). On the other hand, for more common species (*i.e.* with higher mean counts), count data are more likely to show short-term declines. With greater statistical power compared to detection/nondetection data, count data can also describe metapopulation dynamics (Donald and Fuller 1998), or widespread, moderate declines (Joseph and Possingham 2008) that might otherwise be missed. Consequently, in spite of the reduced efficiency for rare species, the more useful single design for omnibus studies concerned with population changes is usually to collect count data.

Conversely, when a project's goal is to determine 'snapshot' species distributions over complete geographic areas (*e.g.* provinces) – and hence, when population trajectories are not of interest – designers should generally choose the more spatially-efficient detection/nondetection atlassing approach, which can survey more land area for less effort. To capitalize on the information-gathering potential of atlas volunteers, recent Canadian atlasses now also collect optional point-count information alongside the principal detection data (*e.g.* Maritimes Breeding Bird Atlas 2006–2010). To my knowledge, whether such an optional contribution is sufficiently controlled and replicated to be useful has not yet been demonstrated in peer-reviewed publications.

The choice of design from a power and an efficiency standpoint must be carefully made with the study's ultimate goals and resources in mind (Elphick 2008; Francis *et al.* 2009); this includes recognizing what designs are likely able to achieve a useful level of statistical power (*e.g.* Bart *et al.* 2004*a*; Purcell *et al.* 2005; Sauer *et al.* 2005), defining precisely what is a 'useful' level of power in a particular case (Field *et al.* 2004), and determining if there is sufficient funding to follow through (Field *et al.* 2007). However, there can also be nontarget benefits to even poorly-powered or inefficient citizen science surveys such as education (*e.g.* Evans *et al.* 2005; Braschler *et al.* 2010) and breadth of knowledge (*e.g.* Peters 2010) which might make the goals of power and efficiency less-important in the overall context of maximizing our ecological knowledge. In the next section, I present an example of such alternative priorities in long-term ecological monitoring.

2.4.2 'Targeted' vs. 'Surveillance' Monitoring – is Deliberate Inefficiency Always Bad?

An ongoing debate in the literature concerns whether ecological monitoring projects should generally be designed around specific hypotheses, conceptual models, and management objectives – for instance, determining whether a population decline in a given game species exists, and whether it is consistent with one of several conceptual models of how that species' reproduction might respond to hunting pressures – or whether less-directed monitoring – for instance an omnibus bird abundance survey without a particular target species and population change mechanism in mind – can be a similarly-useful option. Whereas the former, 'targeted' monitoring strategy (sensu Nichols and Williams 2006) is a more efficient approach for a specific research goal, 'surveillance' monitoring may have other advantages.

On the one hand, the traditional 'surveillance' approach found in older studies (*e.g.* the BBS) has been to record a variety of data that can be detected consistently from year to year, without necessarily having detailed mechanistic hypotheses under evaluation. Proponents of this approach argue that the resulting baseline datasets are a necessary precursor to finer-scaled, more efficient studies in that they draw attention to unusual patterns which might otherwise go unnoticed (Duarte *et al.* 1992; Nisbet 2007; Dickinson *et al.* 2010; Magurran *et al.* 2010; Peters 2010). Long term population indices collected in this way can also qualify steep, short-term population declines as being within a normal range for a particular species (Dunn 2002), and so can prevent unnecessary management responses. However, critics of this approach (*e.g.* Yoccoz *et al.* 2001; Nichols and Williams 2006; Field *et al.* 2007) point out that its less-directed nature can make some data records irrelevant (*e.g.* unused anecdotal records), while other broadly- or sparsely-collected data may lack sufficient statistical power to detect all but the most dramatic changes.

Recognizing that surveillance monitoring can never be as efficient as more-targeted strategies for a given research objective, should we thus consider them to be generally inferior substitutes? In making a case for the superiority of targeted studies, Nichols and Williams (2006) frame the issue philosophically, arguing that, in addition to the lower cost-efficiency, monitoring without mechanistic hypotheses is inconsistent with the conduct of traditional science (*e.g.* Platt 1964), and that the 'rate-of-learning' arising from such an approach is low. Yoccoz *et al.* (2001) share similar views, pointing out that these studies' unavoidable lack of experimental design elements, for instance the lack of randomly-allocated manipulations, limit the ability of such programmes to make strong inferences about more detailed scientific questions, in other words, to suggest causation underlying observed trends (see also Peters 2010).

However, these arguments ignore the value of having a reliable, long-term baseline from which to form mechanistic hypotheses and to design experiments in the first place. For example, our realization that global atmospheric carbon dioxide levels were consistently increasing occurred thanks to decades of background monitoring. Without this dataset that stretched back to 1957, the beginnings of our scientific investigations into the human role in climate change, along with corresponding hypothesis-driven discoveries and conservation initiatives, might have been delayed by decades (Nisbet 2007). Lindenmayer *et al.* (2010) similarly promote long-term monitoring for its ability to detect ecological 'surprises' – unexpected, but important ecological findings that change our understanding of ecosystem function. Such 'surprises' are important precursors to more thorough investigations and corresponding high-quality 'discoveries' that can logically follow. In this sense, 'baseline' surveillance monitoring has a valuable place in long-term, broad-scale ecology as part of a broader group of monitoring *programmes*, which include both baseline and subsequent, targeted studies.

Unfortunately, public support and appreciation for this kind of science can be low. Nichols and Williams (2006) argue that, for wildlife conservation, in spite of there being some initial value in baseline monitoring, there is too little funding available to justify most ongoing surveillance. Accordingly, Nisbet (2007) called long-term monitoring 'Cinderella science' in that it is largely unloved, with a dormant public profile. Duarte *et al.* (1992) suggest that disdain for such long-term monitoring programmes may also occur at the governmental level because they do not typically provide benefits consistent with the timeframe of political appointments. However, the simple and consistent approach of surveillance monitoring is well-suited for citizen science applications in that it can also provide important educational and recreational opportunities in parallel (*e.g.* Trumbull *et al.* 2000; Evans *et al.* 2005; Sullivan *et al.* 2009; Braschler *et al.* 2010). In this way, the scientific inefficiencies of long-term surveillance monitoring can be offset by non-scientific spinoff benefits.

2.4.3 The Ideal Long-Term Monitoring Programme

The best long-term monitoring programmes are probably composed of multiple complementary baseline surveillance monitoring studies and targeted, hypothesis-driven studies arising that reflect a conceptual model of the system of interest (Lindenmayer and Likens 2010; Peters 2010). Although such complementarity might be inefficient in terms of data collection costs, these study programmes leave a manager far better equipped to understand ecological processes (Lindenmayer *et al.* 2010), and hence may also save money in the long term as new management questions present themselves.

Broad, complementary research of this kind has led to a variety of high-quality inferences in practise. For instance, through a combination of baseline and targeted studies of several physical and biological variables, the Hubbard Brook Ecosystem Study thoroughly demonstrated the mechanism and effects of acid precipitation (Lindenmayer *et al.* 2010). Monitoring related ecological variables across multiple, complementary baseline surveys can also assist our recognizing real patterns in noisy biological systems (Lepage and Francis 2002; Francis *et al.* 2005). For instance, Cooper *et al.* (2007) used three separate monitoring datasets to study House Sparrow (*Passer domesticus*) population declines. Each dataset independently implicated interspecific competition as an important contributing factor, allowing the authors to confidently conclude that interspecific competition was the principal culprit. In sum, the initial inefficiencies of using baseline surveillance monitoring might often be justified by the greater confidence of subsequent ecological insights.

To maintain the relevance of baseline surveillance programmes, Lindenmayer and Likens (2010) further argue that we should adopt an 'adaptive monitoring' philosophy which supports design and protocol changes as they are justified by incoming results, new interests, and by emerging technologies, so long as the long-term 'monitoring' results are not compromised. Managers of the CBC and BBS demonstrate this hybrid philosophy in their efforts to continuously improve data collection and modeling (*e.g.*

Francis *et al.* 2004; Sauer *et al.* 2005), and in their support of corresponding research (see examples in Dickinson *et al.* 2010). Regular reviews of survey methods and objectives are important for maintaining the relevance and scientific credibility of any long-term effort (*e.g.* Nichols and Williams 2006; Lindenmayer and Likens 2010; Magurran *et al.* 2010).

Compared to targeted monitoring studies, which have clearer objectives and expected outcomes, a less-focused baseline surveillance project's chances of success (in terms of it providing important biological inferences) increases with its geographic and taxonomic scope. Among other reasons, this is because the results of larger surveillance programmes can be more easily paired with other, separately-funded datasets (*e.g.* weather station time series, geospatial data, fisheries landings) as part of a comprehensive analysis. This in turn supports a greater diversity of specialized research questions. However, such an advantage comes at an important financial cost. Even with volunteer monitors, who are less expensive per datum than dedicated professional scientists, monitoring projects are still costly given their typical scale (Braschler *et al.* 2010). Once initially funded, managers thus need to be highly responsive to perceived problems in their design and analysis in order to provide meaningful, desirable data within their funding limits and timeframe (*e.g.* Field *et al.* 2007). Providing data of such usefulness is not an easy task: Duarte *et al.* (1992) make the frustrating claim that, in marine science,"long-term [*e.g.* surveillance] monitoring programmes are, paradoxically, among the shortest projects. . . many are initiated, but few survive a decade". Field *et al.* (2007) also allude to numerous similar failures with conservation monitoring in Australia.

In sum, opposition to baseline surveillance monitoring compared to more-targeted studies should be directed at solitary projects which are too-narrow in scope, inflexible in management, and underfunded in their implementation. Recognizing the high threshold for success of such projects, it follows that where surveillance is not likely to be useful or possible, for instance because the management action is clear or the funding is not available, it should not occur (McDonald-Madden *et al.* 2010), or if monitoring is no longer necessary, that it stop (Field *et al.* 2004, 2007). When these problems are avoided, surveillance monitoring can thus be an extremely high-quality ecological reference point. To that end, carefully-managed surveillance and targeted monitoring, which share an important interconnectedness, each have their place in ecology and in citizen science.

2.5 Citizen Science and the Future of Avian Monitoring

As we move increasingly to urban living as a global population (Montgomery 2008), and as our growing technological abilities create global-scale ecological impacts, the need to preserve our appreciation for and understanding of natural landscapes is a worldwide concern. Citizen science projects are able to contribute to both of these objectives, engaging a large population of citizens in recognizing their local ecology while providing broad-scale, long-term information to contribute to our knowledge and management of global ecology.

With a long tradition of interest from amateurs, and as an important component of global ecosystems, birds are a relevant focal point for citizen science. Many bird species have already benefited from citizen science projects, the data from which have helped to optimize their population management (*e.g.* Greenwood 2007). Current educational and scientific successes seen are likely to continue, as long as studies are carefully designed and modeled to incorporate uncertainty, with a willingness among managers to modify and expand protocols as new information suggests and permits. Furthermore, the active tradition of expert and non-expert ‘amateurs’ in field ornithology promises continued support for citizen science projects, especially if project protocols are able to accommodate a variety of volunteer lifestyles. As the number and size of our long-term datasets continues to grow, we may find both depressing and unexpected patterns in bird population and other ecological systems. However, armed with this reliable and increasingly extensive knowledge, we will also find ourselves in possession of increasingly powerful and less-deniable evidence for the nature and sometimes, the mechanisms underlying these changes. If collected and used responsibly, citizen science can thus benefit ecosystems worldwide.

Chapter 3

Observer Effects and Avian Call Count Survey Quality: Rare-Species Biases and Overconfidence

3.1 Abstract

Wildlife monitoring surveys are prone to nondetection errors and false positives. To determine factors that affect the incidence of these errors, we built an internet-based survey that simulated avian point counts, and measured error rates among volunteer observers. Using similar-sounding vocalizations from paired rare and common bird species, we measured the effects of species rarity and observer skill, and the influence of a reward system that explicitly encouraged the detection of rare species. Higher self-reported skill levels and common species independently predicted fewer nondetections (probability range: 0.11 [experts, common species] to 0.54 [moderates, rare species]). Overall proportions of detections that were false positives increased significantly as skill level declined (range: 0.06 [experts, common species] to 0.22 [moderates, rare species]). Moderately-skilled observers were significantly more likely to report false-positive records of common species than of rare species, whereas experts were significantly more likely to report false-positives of rare species than of common species. The reward for correctly detecting rare species did not significantly affect these patterns. Because false positives can also result from observers overestimating their own abilities ('overconfidence'), we lastly tested whether observers' beliefs that they had recorded error-free data ('confidence') tended to be incorrect ('overconfident'), and whether this pattern varied with skill. Observer confidence increased significantly with observer skill, whereas overconfidence was uniformly high (overall mean proportion = 0.73). Our results emphasize the value of controlling for observer skill in data collection and modeling and do not support the use of opinion-based (*i.e.* subjective) indications of observer confidence.

3.2 Introduction

Broad-scale and long-term ecological datasets collected by volunteers form an increasingly important component of contemporary wildlife management (Silvertown 2009). Among their many uses, these datasets monitor populations of birds (Link and Sauer 1998; Kéry and Schmid 2006; Julliard *et al.* 2006; Hewson *et al.* 2007), anurans (North American Amphibian Monitoring Program 2011; de Solla *et al.* 2005; Lotz and Allen 2007), invertebrates (Kremen *et al.* 2011; Maritimes Butterfly Atlas 2011), and many marine organisms (*e.g.* Goffredo *et al.* 2010; Ward-Paige *et al.* 2010). Each survey typically records point count and/or detection/nondetection data from a given location over a known time interval, providing broad spatial and temporal data coverage at a minimal cost.

Among surveys of birds and anurans, a substantial proportion of detections are made by ear, without visual confirmation of a species' identity (Dawson and Efford 2009). Unfortunately, accurate auditory identifications can be difficult because many species sound alike (*e.g.* Robbins and Stallcup 1981; McClintock *et al.* 2010*a*). In field settings, different habitats and background noises also affect detection probability (Pacifici *et al.* 2008). Consequently, data collected by auditory surveys generally incorporate some amount of unavoidable observation error.

In spite of such error, volunteer surveys can be scientifically valuable if analyzed appropriately (*i.e.* if uncontrolled variability in detectability can be reduced to less than that of population variability; Johnson 2008). The need among managers for good-quality, broad-scale, long-term ecological data is increasing because of recent and ongoing challenges to global ecosystem stability (*e.g.* U.S. North American Bird Conservation Initiative Committee 2010). Hence, developing methods to extract such information from these surveys is a highly topical and active research concern (Elphick 2008). Reducing the influence of observer error is an important component of this research.

Observer-level errors on detection/nondetection surveys can be divided into two main types: nondetections and false positives (Royle and Link 2006). Nondetections occur when a species is present but not recorded, whereas false positives occur when a species is absent but is nonetheless recorded. False positives are more serious errors because they usually result from the misidentification of a species that is actually present; thus, they are often accompanied by concurrent nondetections (Bart 1985). Under most wildlife survey designs (including our own), the absence of a species is not explicitly recorded, hence, we refer to 'nondetections' instead of 'false negatives', because the latter term implies a declaration-of-absence.

The problem of incomplete detection (*i.e.* nondetection) in animal surveys has been studied for decades, particularly in the avian literature (*e.g.* Bart 1985; Marsden 1999) and direct estimation of corresponding probabilities is becoming routine (*e.g.* Diefenbach *et al.* 2003; Pellet and Schmidt 2005; Etterson *et al.* 2009, but see Rosenstock *et al.* 2002; Johnson 2008). False positive probabilities, on the other hand, are typically assumed to be negligible (*e.g.* MacKenzie *et al.* 2009) and, therefore have received less attention. There is growing evidence from studies of anurans and birds, however, that the frequency of false positives in auditory surveys can be appreciable. For instance, a controlled experiment measuring frog and toad call recognition errors found that 5% of all detection records were incorrect (McClintock *et al.* 2010*a*). Similarly, a study that modeled the occupancy of five bird species using repeated field visits along a North American Breeding Bird Survey (BBS) route estimated false-positive probabilities per detection event of up to 0.165 (Royle and Link 2006). Four other sets of controlled birdsong simulations showed that false-positives comprised 1–14% of the total number of detections (mean = 5.8%; Bart 1985; Simons *et al.* 2007; Alldredge *et al.* 2008; Campbell and Francis 2011). Mathematical simulations have shown that failing to account for false positives of these magnitudes can lead to substantially biased estimates of species occupancy parameters (Royle and Link 2006; McClintock *et al.* 2010*b*; Miller *et al.* 2011).

At present, there are limited practical opportunities to correct for both false positives and nondetections simultaneously. Current published approaches that make such corrections ('misclassification models') require data from replicated surveys (*e.g.* multiple visits made during the same season; Royle and Link 2006; McClintock *et al.* 2010*b*; Miller *et al.* 2011). Unfortunately, without some indication of the reliability of the observation (Miller *et al.* 2011), misclassification modeling may yield biased occupancy estimators in the presence of varying levels of observer skill (Fitzpatrick *et al.* 2009) and when error rates are not consistent among sites (McClintock *et al.* 2010*b*). By design, they are also not suitable for surveys that lack replicated data (*e.g.* most BBS routes, which are surveyed once annually). Collectively, most current study designs and modeling approaches therefore have a limited ability to address important detection errors.

One approach to reduce the influence of detection errors is to address factors contributing to their occurrence (Raitt 1981; Johnson 2008; McClintock *et al.* 2010*a*). We propose that an observer's 'state-of-mind', which we define here as being the sum of conscious and unconscious biases that can affect decision-making behaviour (*e.g.* Croskerry 2002; Lane *et al.* 2007), might constitute such a factor. Although previous authors have speculated that an observer's "attitude" (Faanes and Bystrak 1981), "carelessness" (Robbins and Stallcup 1981), and preferences and expectations (Balph and Balph 1983) might lead to identification errors on call count surveys, to our knowledge, there has been little quantitative research addressing this overall theme in ecology.

Nonetheless, such sources of error could be quite important. For instance, bird-watchers – and possibly surveyors of other taxa – are often motivated to detect and report the presence of rare species (Sullivan *et al.* 2009), and we hypothesize that such a preference might bias an observer to both detect more rare species under ambiguous circumstances ('observer expectancy effects'; Miller and Turnbull 1986; Lane *et al.* 2007) and similarly, to be more attentive to the sounds of rare species ('search-image' detection biases; Callahan *et al.* 2003). These biases might lead to correspondingly fewer nondetections and more false positives for rare species than for common species. On the other hand, rarer species could instead be prone to more nondetections than common species if an observer arbitrarily rules out the possibility of a given rare species being present at all, on the basis of its rarity (the 'playing the odds' bias; Croskerry 2002).

Exploring this theme, Bart (1985) re-analyzed an experimental call-count survey dataset, in part to determine whether observers tend to detect particular species more often than others. He indeed found that detection error rates varied among species; however his focus was not on the detection of rare versus common species specifically. Two recent studies have shown that detection error rates do vary among rare and common species: species that call less often on field recordings of bird choruses tend to be associated with greater numbers of detection errors than frequently calling species (Rempel *et al.* 2005; Campbell and Francis 2011). However, those studies did not test for mechanisms underlying this pattern, for instance whether these errors tend to arise from a lack of observer knowledge, and/or confusion with common species. Further research is thus needed that specifically controls for the effects of observer skill and the potential for rare and common species to sound alike.

Along with biases for or against the detection of rare species, unfounded observer confidence ('overconfidence'; Moore and Healy 2008) in a particular species identification might also be an important source of detection errors. An overconfident observer tends to overestimate his or her performance on a given task, and thus is more prone to making detection errors than less overconfident observers, all else being equal. Given that overconfidence tends to occur more commonly among self-assessed experts than among novices (Larrick *et al.* 2007), and that many call count surveys involve expert volunteers (*e.g.* Sauer *et al.* 1994; Genet and Sargent 2003), overconfidence might explain a number of false-positive errors in survey datasets. However, we are not aware of research that has quantified its prevalence in this ecological context.

We used an internet-based survey that mimicked an avian field point count to address these knowledge gaps. We determined rates of nondetections, false positives and overconfidence among a set of volunteer observers to determine (i) whether observers of varying skill levels are more or less prone to detect rare species more or less often than similar-sounding common species; (ii) whether an explicit incentive to correctly detect rare species affects error rates; and lastly, (iii) whether overconfidence is common among observers of different skill levels.

3.3 Methods

We created an internet-based survey designed to mimic what observers might hear during an avian point count. The survey was composed of 16 simulated bird choruses ('scenarios') of known species, each lasting 30 s. We recruited volunteer observers to participate in the survey using e-mails sent to rare-bird and natural-history e-mail listservers in the Maritimes provinces, Canada (n = 3 listservers) and the northeastern United States (hereafter "New England"; n = 3 listservers; USA; see Acknowledgments), and by word-of-mouth. Upon visiting the survey web site, observers were first presented with an introductory page asking that they have a basic familiarity with the vocalizations of 38 candidate bird species, which we indicated might be presented in the survey. Only 12 species were actually used. We provided hyperlinks to examples of each candidate species' vocalizations. Observers were told that the featured choruses typified birds found in mixed or predominantly coniferous forest habitats (including wet brush) of eastern North America, but they were otherwise given no further information about the structure or contents of the testing scenarios.

Following this initial screening, observers were asked to declare their skill level from a list of five options (No Experience, Beginner, Moderate, Advanced, Expert) that were not defined further. They were then asked to listen to each of the 16 scenarios once, manually beginning playback of each new scenario when ready, and then to indicate which birds were heard in each scenario using only the checklist of 38 candidate species. Re-playing the scenario was possible, but explicitly discouraged. Observers were not asked to count the number of individuals calling. Finally, to gauge their confidence and test for overconfidence, the survey asked observers to indicate at the end of each scenario whether they believed that they had correctly identified all species that were present.

We created all scenarios using audio samples of vocalizations (*i.e.* calls and songs) collected with permission from the Macaulay Library of the Cornell Laboratory of Ornithology (http://macaulaylibrary.org) and modified to remove background noises and normalize volume levels using the free audio manipulation software Audacity (http://audacity.sourceforge.net). Each scenario featured the vocalizations of 6 species, sampled with replacement from a pool of 12 species (consisting of 6 similar-sounding species pairs of opposing rarities; Table 3.1). With the exception of the Black-capped and Boreal chickadees (for which we used *chick-a-dee*-type calls), all vocalizations used in the scenarios were songs. Vocalizations ranged in length from 0.8 s (Alder Flycatcher) to 2.5 s (Song Sparrow) and were repeated three times per species, arranged arbitrarily within the scenarios.

We overlapped the transitions between ∼90% of successive vocalizations to make scenarios comparable to a natural field situation. The maximum length of time between the remaining nonoverlapping vocalizations was ∼1 s. To add standardized

Table 3.1. Species used in the survey scenario recordings, grouped by species pairs (A-F) and rarity classes (Common or Rare), assigned according to the number of Maritimes Breeding Bird Atlas squares in which each species was present (percentage of 1499 possible squares in parentheses).

Species Pair	Common Name	Scientific Name	Rarity (Percent MBBA squares)
A	Alder Flycatcher	*Empidonax alnorum*	Common (65.7)
A	Olive-sided Flycatcher	*Contopus cooperi*	Rare (30.2)
B	American Robin	*Turdus migratorius*	Common (84.5)
B	Rose-breasted Grosbeak	*Pheucticus ludovicianus*	Rare (26.6)
C	Black-capped Chickadee	*Poecile atricapillus*	Common (77.6)
C	Boreal Chickadee	*P. hudsonicus*	Rare (38.8)
D	Dark-eyed Junco	*Junco hyemalis*	Common (70.2)
D	Palm Warbler	*Setophaga palmarum*	Rare (37.6)
E	Swainson's Thrush	*Catharus ustulatus*	Common (60.4)
E	Veery	*C. fuscescens*	Rare (37.0; M only)[a]
F	Song Sparrow	*Melospiza melodia*	Common (73.2)
F	Lincoln's Sparrow	*M. lincolnii*	Rare (27.0)

[a] "M only" indicates species that are relatively rare in the Maritimes but common in New England, and thus were scored as "Common" for New England survey results

natural background noise to each scenario, we also superimposed a sequence of ambient cricket noises taken from a Macaulay audio sample on the sequence of bird vocalizations (maximum cricket amplitude [dB] was $<1\%$ of peak birdsong amplitude). The loudness of each vocalization was consistent among all species.

Our pool of 12 potential species was composed of 6 species pairs, each of which shared qualitatively similar vocalizations (*e.g.* American Robin and Rose-breasted Grosbeak; Robbins and Stallcup 1981, Table 3.1). Each member of a species pair was classified as a 'rare' or 'common' variant according to the extent of its range in the Maritimes Breeding Bird Atlas (2006–2010). Specifically, we used the number of 10 km × 10 km 'atlas squares' in which a species was present to determine relative rarity, with the rare variant of the pair occurring in fewer squares than the common variant (Table 3.1). We changed one rarity classification from "rare" to "common" for Veerys detected by New England observers ($n = 30/52$ observers). This change reflected its increased density in this more southern region (Bevier *et al.* 2005). Thus, in these cases, one of the six species pairs contained only common variants. Because detection events were not modeled as explicit choices between rare and common variant pairs, but instead among 'rare' and 'common' species collectively (see below), we did not expect this change to affect the quality of our inferences. In addition, there was no significant difference between the distribution of skill levels among Maritimes and New England observers which might otherwise bias detection data ($\chi^2_3 = 0.275$, $P = 0.965$)

One member of each species pair was randomly assigned to half of the scenarios; the second half of the scenarios featured the other member. In this way, no two members of a species pair appeared together simultaneously. Hence, false positives involving the species pairs could largely be interpreted as mistakes for the rarer or for the common variant. All scenarios had six distinct vocalizations (representing one member of each of the six species pairs), repeated three times each (Table 3.2). We duplicated each scenario and alternated the duplicates randomly alongside the originals, for a total of 16 scenarios presented to each observer (Table 3.2). We informed observers that every second scenario would be 'scored', and that correctly detecting rarer species was worth more points than correctly detecting common species. We then posted and regularly updated the top five high scores alongside user ID codes

Table 3.2. Summary of experimental design of our internet-based survey. "Scenarios" are the separate audio tracks played sequentially to the observers.

Item	n
Scenarios	16 (8 unique × 2 for incentive treatment)
Total Species	12 (6 vocalization group pairs × 2)
Number of scenarios in which a given species is present	8 (4 × 2 for incentive treatment)
Species vocalizing per scenario	6
Discrete vocalizations per species, per scenario	3
Discrete vocalizations per scenario (all species)	18

on the survey website. Our intent was to create and measure the effect of an explicit incentive to detect rare species on detection error rates. Observers were not told that the scenarios were duplicated, and we assumed that the scenarios were too similar-sounding and complex to be recognized as such. We did not expect this randomized, alternating design to show any important learning-effects biases; nonetheless, we controlled for any such systematic differences between earlier and later scenarios (see below).

We traded off the statistical need to present a large number of scenario replicates to our observers against the need to present realistic (longer) survey lengths. Our survey length of 30 s was substantially shorter than that of typical roadside point counts, which tend to last for 3 to 5 minutes, but roadside anuran survey research has shown that most species detections occur within the first 60 s (Shirose *et al.* 1997). Also, the species richness we presented was small per scenario ($n = 6$), and thus arguably manageable under these constraints. Hence, we assume that the challenge posed to our volunteer observers was appreciable, but not unreasonable.

Modeling details. — We defined a correct detection as occurring when a species that was present in a scenario was reported as such, and false-positive detection as occurring when a species that was not present in a scenario was similarly reported as being present. The probability of making a correct detection for a given species is equivalent to 1 minus the probability of making a nondetection error; here, we modeled correct detections in place of nondetections because the conceptual interpretation is more intuitive. Rates of correct (and non-) detections vary independently of false positives.

To determine the effect of species rarity on the incidence of false positives, we first recognized that many 'phantom' species (sensu Bart and Schoultz 1984; McClintock *et al.* 2010*b*) that did not appear in any scenario were nonetheless identified repeatedly from the survey's list of 38 candidate species (Table 3.3). As with the 'playback' (*i.e.* not-phantom) species pairs, we defined each phantom species as being either rare or common so that their data records could be modeled. Here, we determined the relative rarity values for each phantom species again using the Maritimes Breeding Bird Atlas detection records (2006–2010), but using the percentage of atlas squares occupied by the most abundant of the 'rare' playback species (38.8%) as the threshold value distinguishing 'rare' phantom species from 'common' phantom species (Table 3.3). The maximum percentage of atlas squares occupied by 'rare' phantom species was 18.3%; the minimum percentage of atlas squares occupied by a 'common' phantom species was 61.4% (Table 3.3).

One phantom species (Eastern Phoebe) was common in the northeastern United States compared to most Maritimes regions (Weeks Jr. 1994). Hence, we scored its detection records as 'rare' if observations came from Maritimes survey participants and 'common' if they came from New England survey participants. To simplify statistical analyses, we also arbitrarily discarded detection records for phantom species detected < 7 times out of 4025 total detection records (Table 3.3). We also discarded records from the single "Beginner", because there was no replication of this skill level.

We used generalized linear mixed models ('GLMMs') to determine expected probabilities of correct detections and expected proportions of all detections that were false positives. Generalized linear mixed models incorporate random effects structures that recognize group-level deviations from overall patterns (Venables and Ripley 2002). We modeled correct detections and false-positive proportions as binomial responses, and incorporated random effects structures that accounted for differences in error rates among observers (both models) and species pairs (correct detection model only). Our choice to model false positives as proportions of all detections

Table 3.3. Species that were candidates for detection but not included in the scenario recordings ('phantom' species). Rarity classes (Common or Rare) were assigned to those species that were reported at least 7 times among 4025 species records ($n = 19$ records or $< 0.5\%$ of the total). Rarity classes were assigned according to the number of Maritimes Breeding Bird Atlas squares in which each species was present (percentage of 1499 possible squares in parentheses); here, 'rare' species were found in $< 38.8\%$ of atlas squares.

Common Name	Scientific Name	Rarity (Percent MBBA squares)[a]
American Woodcock	*Scolopax minor*	
Barred Owl	*Strix varia*	
Belted Kingfisher	*Ceryle alcyon*	
Black-and-white Warbler	*Mniotilta varia*	Common (61.4)
Black-throated Green Warbler	*Setophaga virens*	
Common Grackle	*Quiscalus quiscula*	
Common Nighthawk	*Chordeiles minor*	
Common Yellowthroat	*Geothlypis trichas*	
Eastern Phoebe	*Sayornis phoebe*	Rare (18.3; M only)[b]
Eastern Towhee	*Pipilo erythrophthalmus*	
European Starling	*Sturnus vulgaris*	
Fox Sparrow	*Passerella iliaca*	Rare (11.1)
Great Horned Owl	*Bubo virginianus*	
Hairy Woodpecker	*Picoides villosus*	
Hermit Thrush	*Catharus guttatus*	Common (72.4)
Ovenbird	*Seiurus aurocapilla*	
Pine Warbler	*S. pinus*	Rare (4.8)
Red-eyed Vireo	*Vireo olivaceus*	Common (74.6)
Red-tailed Hawk	*Buteo jamaicensis*	
Rock Pigeon	*Columba livia*	
Scarlet Tanager	*Piranga olivacea*	Rare (6.2)
Eastern Whip-poor-will	*Caprimulgus vociferus*	
Willow Flycatcher	*Empidonax traillii*	Rare (1.9)
Wilson's Warbler	*Cardellina pusilla*	Rare (14.3)
Yellow Warbler	*S. petechia*	
Yellow-rumped Warbler	*S. coronata*	Common (73.4)

[a] Rarity and percent MBBA square values were calculated only for those phantom species detected 7 times or more among all observers and scenarios, because only data from these phantom species were included in the predictive models.

[b] "M only" indicates species which are relatively rare in the Maritimes but common in New England and thus scored as "Common" for New England survey results

is consistent with previous studies (*e.g.* Simons *et al.* 2007; Alldredge *et al.* 2008; McClintock *et al.* 2010*a*).

Each of the models of correct detections and false positives allowed us to estimate error rates while measuring the influence of several predictors. We used the GLMMs to model (1) how the rates of each type of error varied among rare and common species; (2) the effect of the incentive treatment rewarding the correct detection of rare species over common ones; (3) how observer skill was related to error rates; and (4) any skill- and incentive-dependent differences (interactions) in the detection of species of each rarity class. To correct for skill-dependent changes in observer ability over the course of the survey (*e.g.* learning, changes in interest level), we also included (5) the chronological scenario number and its interaction with observer skill as additional covariates.

To estimate the probability of making a correct detection for a given species rarity class and scenario, we built a dataset consisting of a record for each correct detection (1 = the species was present and detected) and each nondetection (0 = the species was present but not detected), and excluding all false positives. We used a total of 4416 records of correct detections ($n = 2864$) and nondetections ($n = 1552$).

We modeled the expected probability of making a correct detection for a given species on a given scenario as a mixed-effects Bernoulli process using the package *lme4* in R version 2.13.0 (Bates and Maechler 2010; R Development Core Team 2011). In this model, in addition to recognizing differences in correct detection probability among observers as random intercepts, we also recognized variation in mean correct detection probability between species pair-groups, given that some species' calls are more easily detected than others (Alldredge *et al.* 2007*a*):

$$\begin{aligned} logit(P(Y_{ijkl} = 1)) = \beta_0 + \beta_1 \cdot Rarity_i + \beta_2 \cdot Skill_j + \beta_3 \cdot Scenario_k \\ + \beta_4 \cdot Skill_j : Scenario_k + \beta_5 \cdot Skill_j : Rarity_i + \beta_6 \cdot Incentive_k \\ + \beta_7 \cdot Rarity_i : Incentive_k + b_{1_j} + b_{2_l} \end{aligned} \quad (3.1)$$

where $Y_{ijkl} = 1$ when a species is correctly scored as being present; $i = 1$ of 2 rarity classes; $j = 1 \ldots, 52$ observers; $k = 1, \ldots, n_j$ scenarios completed per observer; and $l = 1, \ldots, 6$ species pairs. Random effects b_{1_j} and b_{2_l} are independently and normally distributed intercepts for observers and for species pairs, respectively, with means zero and with standard deviations estimated from the data.

We then calculated the proportion of all detections for each observer-within-scenario that were false positives for each of the rare and common species groups. For instance, if observer A, listening to scenario 1 incorrectly reported the presence of two rare species and one common species, and correctly reported four common species and three rare species, his false-positive proportions would be 0.2 and 0.1 for rare and common species, respectively. In total, we modeled 1429 false positive proportions. We estimated the proportion of false positives per rarity class per scenario as a mixed-effects binomial process with the same predictors as equation 3.1, but here with a simpler random-effects structure, as follows:

$$\begin{aligned} logit(P(Y_{ijk}=1)) &= \beta_0 + \beta_1 \cdot Rarity_i + \beta_2 \cdot Skill_j + \beta_3 \cdot Scenario_k \\ &+ \beta_4 \cdot Skill_j : Scenario_k + \beta_5 \cdot Skill_j : Rarity_i + \beta_6 \cdot Incentive_k \\ &+ \beta_7 \cdot Rarity_i : Incentive_k + b_{1_j} \end{aligned} \quad (3.2)$$

where Y_{ijk} is the proportion of all detections that were incorrect (false-positive) for a given observer, scenario and species rarity class; $i = 1$ of 2 rarity classes; $j = 1 \ldots, 52$ observers; $k = 1, \ldots, n_j$ scenarios completed per observer; and b_{1_j} is a normally distributed random intercept for observers with mean zero and standard deviation estimated from the data.

To measure and model confidence levels among survey participants, we first asked observers at the end of each scenario if they believed that they had correctly accounted for all species present. If they answered ‘yes’, we considered that scenario and its responses to be ‘confident’. We then calculated the proportion of scenarios completed by each observer that were confident.

We also calculated the proportion of overconfident scenarios. We defined an overconfident scenario as one in which an observer made at least one detection error while also declaring confidence. This measure thus indicated the probability that a given observer’s declaration of confidence was incorrect.

Using generalized linear models (GLMs), we modeled both the incidence of declared confidence, and the incidence of overconfidence as functions of observer skill.

Our confidence data were collected at a different resolution than our detection data; here, each observer contributed one confidence record per scenario. Accordingly, we built the following models:

$$logit(P(Y_{1\cdot ij} = 1)) = \beta_0 + \beta_1 \cdot Skill_i \tag{3.3}$$

$$logit(P(Y_{2\cdot ik} = 1)) = \beta_0 + \beta_1 \cdot Skill_i \tag{3.4}$$

where $Y_{1\cdot ij} = 1$ occurs when a participant declares that a particular survey scenario was scored entirely correctly ('declared confidence') and $Y_{2\cdot ik} = 1$ occurs when a declaration of confidence is incorrect ('overconfidence'); $i = 1 \ldots, 52$ observers; $j = 1, \ldots, n_i$ scenarios completed per observer; and $k = 1, \ldots, m_i$ confident scenarios per observer.

All models were checked for fit quality by examining conventional or binned residual plots (Gelman and Hill 2007), and results were compared visually with plotted raw data to check for consistency. Unless otherwise specified, results are presented as means ± SD.

3.4 Results

We modeled data from observers representing three self-reported skill levels: "Moderate" ($n = 17$), "Advanced" ($n = 26$) and "Expert" ($n = 9$), from the Canadian provinces of New Brunswick, Nova Scotia and Prince Edward Island and the New England states of Maine, New Hampshire and Vermont.

Most observers (80.8%) completed all 16 scenarios (mean number of scenarios completed = 14.15 ± 3.31). We suspect that those who failed to complete all 16 scenarios largely did so in error, rather than out of fatigue or disinterest. This is because the survey was composed of four webpages containing four scenarios each, and most of the missed scenarios were in groups of four sequential scenarios located on the same web page.

Observers with higher skill levels were significantly more likely to correctly detect any given species than observers of lower skill levels (Figure 3.1A and Table 3.4). Across all skill levels, all observers were also equally and significantly less likely to

Table 3.4. Factors that affected the probability of correctly detecting a species on a given scenario (Equation 3.1; $n = 4416$ correct and nondetections distributed among 52 observers). In this binomial model, σ_{b_1}, the standard deviation about the observer random effects, is 0.72 and σ_{b_2}, the standard deviation about the species-pair random effects, is 1.02. All values are on the logit scale.

Factor	Estimate	SE	z value	P
(Intercept)	-0.516	0.547	-0.943	0.345
Rarity	**-0.576**	**0.228**	**-2.531**	**0.011**
Skill	**0.820**	**0.198**	**4.143**	**<0.001**
Scenario	0.017	0.023	0.727	0.467
Incentive	-0.046	0.102	-0.454	0.650
Skill:scenario	0.001	0.012	0.115	0.908
Rarity:skill	-0.039	0.115	-0.337	0.736
Rarity:incentive	-0.045	0.147	-0.307	0.759

correctly detect rare species than common ones (Figure 3.1A and Table 3.4). Neither the incentive nor the scenario number was significantly related to correct detection rates among and within observers and skill levels (Table 3.4).

The expected proportion of species correctly detected per scenario for each skill level ranged from 0.61 (95% CI: 0.40-0.79; Moderate) to 0.89 (95% CI: 0.77-0.95; Expert) for common species, and from 0.46 (95% CI: 0.27-0.67; Moderate) to 0.81 (95% CI: 0.63-0.91; Expert) for rare species (Figure 3.1A). Subtracting these values from 1.0 gives a set of nondetection probabilities that range from 0.11 (Expert skill, common species) to 0.54 (Moderate skill, rare species).

Summed across both species rarity groups, the proportion of false positives declined significantly with increasing skill level (Table 3.5). However, skill level also interacted significantly with species rarity. Here, moderately-skilled observers falsely detected common species more often than rare species, whereas experts falsely detected rare species more often than common ones (Figure 3.1B and Table 3.5). Again, neither the incentive nor the scenario number was significantly related to the occurrence of false positives across or within skill levels (Table 3.5).

The expected proportion of false positives per scenario for each skill level ranged from 0.061 (95% CI: 0.043-0.085; Expert) to 0.218 (95% CI: 0.170-0.280; Moderate) for common species, and from 0.119 (95% CI: 0.087-0.164; Expert) to 0.120 (95%

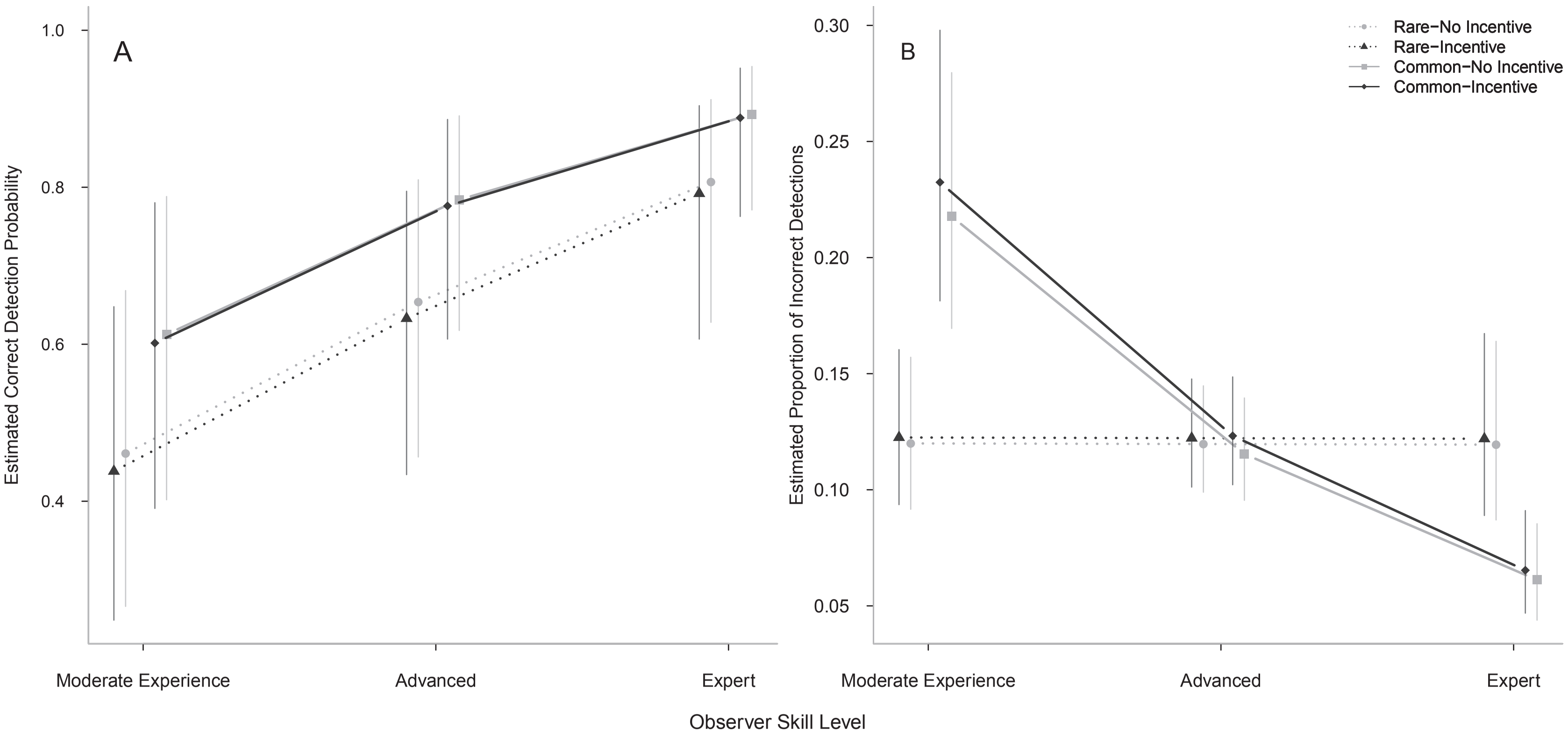

Figure 3.1. Summary of (A) the predicted probability of correctly detecting a species and (B) the predicted proportion of false positives per scenario, grouped by both species rarity and whether the incentive to detect rare species was in effect, ordered by self-reported skill level. Error bars are 95% confidence intervals.

Table 3.5. Factors that affected the number of false positives for a given scenario and species rarity group (Equation 3.2, $n = 1424$ counts of false positives distributed among 52 observers). In this binomial model, σ_{b_1}, the standard deviation about the observer random effects, is 0.41. All values are on the logit scale.

Factor	Estimate	SE	t value	P
(Intercept)	**-0.893**	**0.287**	**-3.114**	**0.002**
Rarity	**-1.228**	**0.224**	**-5.473**	**<0.001**
Skill	**-0.598**	**0.149**	**-3.997**	**<0.001**
Scenario	0.0003	0.022	0.016	0.987
Incentive	0.065	0.101	0.651	0.515
Skill:scenario	-0.004	0.012	-0.370	0.711
Rarity:skill	**0.632**	**0.112**	**5.649**	**<0.001**
Rarity:incentive	-0.044	0.147	-0.302	0.762

CI: 0.091-0.157; Moderate) for rare species (Figure 3.1B).

A tabular summary of the correct detection and false-positive frequencies, indexed by species and observer skill level, can be found in Table 3.6.

Table 3.6. Summary of correct detection and false-positive data, grouped by species and observer skill level. ‘Correct’ count data are the total number of correct detections for a given species among all observers and scenarios (‘observer-scenarios’). The proportion correct (in parentheses) is the number of correct detections divided by the total number of times that species was played among all observer-scenarios. ‘False positive’ count data are the total number of false positives for a given species among all observer-scenarios. The proportion of false positives per scenario (in parentheses) is the number of false positives divided by the total number of observer-scenarios. This value is different from the modeled proportion of false positives per scenario (results here are summarized across multiple observers and scenarios).

			Moderate (248 observer-scenarios)		Advanced (372 observer-scenarios)		Expert (116 observer-scenarios)	
Group	Species[a]	Rarity	Correct (Proportion Correct)	False Positive (per Scenario)	Correct (Proportion Correct)	False Positive (per Scenario)	Correct (Proportion Correct)	False-positive (per scenario)
A	ALFL	C	89 (0.7)	17 (0.07)	167 (0.86)	3 (0.01)	53 (0.95)	0 (0)
A	OSFL	R	96 (0.8)	8 (0.03)	171 (0.97)	7 (0.02)	60 (1)	4 (0.03)
B	AMRO	C	87 (0.69)	48 (0.19)	134 (0.71)	55 (0.15)	57 (0.92)	12 (0.1)
B	RBGR	R	55 (0.45)	8 (0.03)	88 (0.48)	9 (0.02)	41 (0.76)	2 (0.02)
C	BCCH	C	116 (0.92)	34 (0.14)	179 (0.95)	35 (0.09)	56 (0.9)	1 (0.01)
C	BOCH	R	52 (0.43)	6 (0.02)	100 (0.55)	2 (0.01)	39 (0.72)	3 (0.03)

[a] See last page of table for abbreviations

[b] Rare only in the Maritimes provinces

Continued on next page

Table 3.6, continued

			Moderate (248 observer-scenarios)		Advanced (372 observer-scenarios)		Expert (116 observer-scenarios)	
Group	Species[a]	Rarity	Correct (Proportion Correct)	False Positive (per Scenario)	Correct (Proportion Correct)	False Positive (per Scenario)	Correct (Proportion Correct)	False-positive (per scenario)
D	DEJU	C	32 (0.25)	28 (0.11)	70 (0.37)	39 (0.1)	26 (0.42)	21 (0.18)
D	PAWA	R	25 (0.2)	14 (0.06)	62 (0.34)	25 (0.07)	21 (0.39)	13 (0.11)
F	LISP	R	50 (0.39)	16 (0.06)	101 (0.53)	36 (0.1)	54 (0.84)	15 (0.13)
F	SOSP	C	53 (0.44)	5 (0.02)	118 (0.65)	6 (0.02)	48 (0.92)	3 (0.03)
E	SWTH	C	63 (0.5)	8 (0.03)	146 (0.76)	20 (0.05)	54 (0.95)	0 (0)
E	VEER	R[b]	89 (0.73)	19 (0.08)	153 (0.85)	16 (0.04)	59 (1)	1 (0.01)
Phantom	BAWW	C		2 (0.01)		1 (0)		0
Phantom	EAPH	R[b]		22 (0.09)		5 (0.01)		0
Phantom	FOSP	R		3 (0.01)		32 (0.09)		1 (0.01)
Phantom	HETH	C		34 (0.14)		12 (0.03)		0
Phantom	PIWA	R		23 (0.09)		34 (0.09)		10 (0.09)
Phantom	REVI	C		4 (0.02)		14 (0.04)		0
Phantom	SCTA	R		14 (0.06)		26 (0.07)		7 (0.06)

[a] See last page of table for abbreviations

[b] Rare only in the Maritimes provinces

Continued on next page

Table 3.6, continued

			Moderate (248 observer-scenarios)		Advanced (372 observer-scenarios)		Expert (116 observer-scenarios)	
Group	Species[a]	Rarity	Correct (Proportion Correct)	False Positive (per Scenario)	Correct (Proportion Correct)	False Positive (per Scenario)	Correct (Proportion Correct)	False-positive (per scenario)
Phantom	WIFL	R		11 (0.04)		14 (0.04)		1 (0.01)
Phantom	WIWA	R		5 (0.02)		23 (0.06)		12 (0.1)
Phantom	YRWA	C		7 (0.03)		13 (0.03)		5 (0.04)

[a] Abbreviations: ALFL = Alder Flycatcher (*Empidonax alnorum*); OSFL= Olive-sided Flycatcher (*Contopus cooperi*); AMRO = American Robin (*Turdus migratorius*); RBGR = Rose-breasted Grosbeak (*Pheucticus ludovicianus*); BCCH = Black-capped Chickadee (*Poecile atricapillus*); BOCH = Boreal Chickadee (*P. hudsonicus*); DEJU = Dark-eyed Junco (*Junco hyemalis*); PAWA = Palm Warbler (*Setophaga palmarum*); SWTH = Swainsons Thrush (*Catharus ustulatus*); VEER = Veery (*C. fuscescens*); SOSP = Song Sparrow (*Melospiza melodia*); LISP = Lincolns Sparrow (*M. lincolnii*), BAWW = Black-and-white Warbler (*Mniotilta varia*); EAPH = Eastern Phoebe (*Sayornia phoebe*); FOSP = Fox Sparrow (*Passerella iliaca*); HETH = Hermit Thrush (*C. guttatus*); PIWA = Pine Warbler (*S. pinus*); REVI = Red-eyed Vireo (*Vireo olivaceus*); SCTA = Scarlet Tanager (*Piranga olivacea*); WIFL = Willow Flycatcher (*E. traillii*); YRWA = Yellow-rumped Warbler (*S. coronata*)

[b] Rare only in the Maritimes provinces

The proportion of scenarios for which an observer declared confidence increased significantly with self-assessed observer skill ($\beta_1 = 1.376 \pm 0.148, P < 0.001$; Equation 3.3 and Figure 3.2A), with model-estimated values increasing from 0.079 (Moderate; 95%CI: 0.06-0.11) to 0.575 (Expert; 95%CI: 0.50-0.65). Among those surveyors who declared confidence on at least one survey scenario ($n = 28$), there was no significant difference in the amount of overconfidence among skill levels ($\beta_1 = -0.366 \pm 0.288, P = 0.204$; Equation 3.4 and Figure 3.2B). Model-estimated proportions of overconfident scenarios (overall mean 0.73) ranged from 0.80 (Moderate; 95%CI: 0.64-0.90) to 0.66 (Expert; 95%CI: 0.54-0.76); this difference was not statistically significant.

3.5 Discussion

We found significant relationships between detection error rates and each of observer skill and species rarity. In our models, the probability of making a nondetection error decreased with observer skill and among common species, as did the proportion of responses that were false positives. A significant interaction between skill and species rarity for false positives also indicated that among moderately-skilled observers, the majority of false positives were of common species, and among experts, the majority of false positives were of rare species. We also found no evidence that an incentive to detect rare species affected error rates. Finally, observers of all skill levels were overconfident, with 73% of scenarios completed by confident observers of any skill level having at least one error. Below, we address each of these findings in turn.

The range of observed nondetection error probabilities (0.11–0.54) is consistent with the results from similar experiments. For instance, Alldredge *et al.* (2007*a*) calculated values ranging from 0.17 to 0.59, depending on the species and singing rate. Similarly, Simons *et al.* (2007) found probabilities ranging from 0.26 to 0.68 overall, and Bart (1985) reported a probability of 0.30 on average. Using unretouched field recordings, Campbell and Francis (2011) also reported a value of 0.23. Contributing to these errors were slower singing rates (Alldredge *et al.* 2007*a*), louder background noise (Simons *et al.* 2007), and increased local species rarity (Campbell and Francis

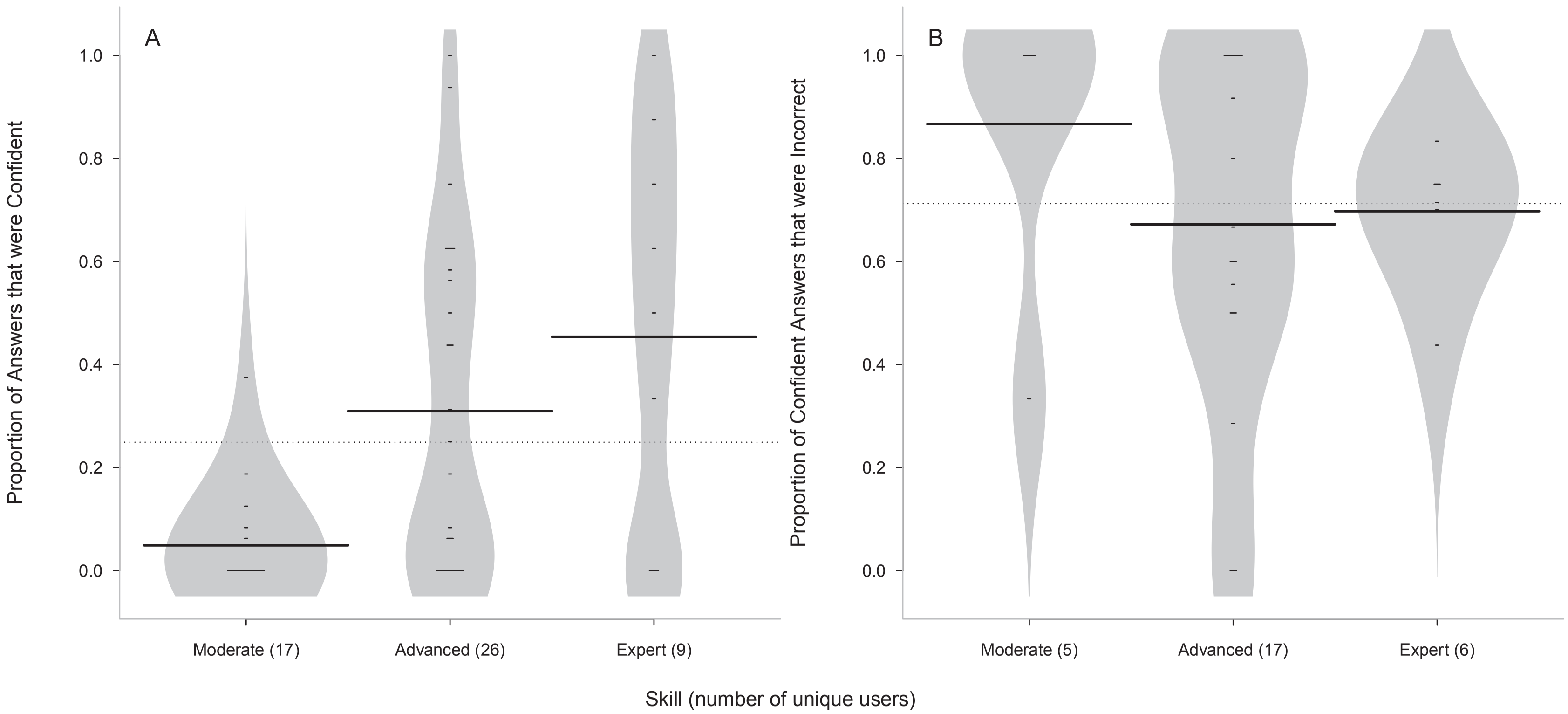

Figure 3.2. Beanplot summary (Kampstra 2008) of the proportion of scenarios that were 'confident' (A) and the proportion of confident scenarios having at least one error ('overconfidence'; B), grouped by self-reported skill level. 'Confident' scenarios were those scenarios in which an observer declared that he or she had correctly detected all species present. In each 'bean', small ticks correspond to individual values for a given skill level, and are scaled by length according to their frequency. The longer solid lines are the mean values for each bean, and the dotted line is the overall mean. Note that these are raw data values; modeled predictions differ only slightly.

2011). Our research shows that, controlling for similarity in vocalizations, increasing rarity at the population scale and decreasing observer skill are also important predictors of species detection.

The observed frequency of false positives (0.06–0.22) was also consistent with past research. For instance, Lotz and Allen (2007) found similar proportions of scenarios that had at least one incorrectly detected anuran (in the absence of similar-sounding equivalents; 0.19 and 0.238, two regions), and Campbell and Francis (2011) found that ∼14% of bird detection records could not be confirmed from simultaneous field recordings, which suggests that they were false positives. Our results were, however, higher than some previously-published rates – Simons *et al.* 2007 (0.01–0.04), Alldredge *et al.* 2008 (0–0.01), and McClintock *et al.* 2010a (0.05) – possibly because we had ambiguous candidate species broadcast over a relatively short period (*i.e.* higher difficulty), and likely a lower average observer skill level. Although Royle and Link (2006) also found the probabilities of detecting a bird species, given its absence (‘p_{10}’) to range from 0.007 to 0.165, this statistic differs from what has been calculated from most studies, including ours (*i.e.* proportion of all detections that are incorrect), and so is not directly comparable. Nonetheless, both our results and this related observation suggest that false-positive rates in avian field detection data are nontrivial.

Several previous studies have shown significant differences among individual observers in their ability to detect and identify animal sounds (*e.g.* Shirose *et al.* 1997; Link and Sauer 1998; McLaren and Cadman 1999; Alldredge *et al.* 2007*a*). However, few have found relationships specifically tied to observer skill, probably because most used a homogeneous group of expert participants who are all highly competent in spite of differences in their amateur or professional status, or in their high absolute levels of experience (*e.g.* Genet and Sargent 2003; Lotz and Allen 2007; McClintock *et al.* 2010*a*). Conversely, our more heterogeneous group demonstrated expected decreases in detection errors with increasing observer skill. Our use of self-assessment of observer skill therefore appeared to successfully capture real differences in ability; this suggests that self-assessment can be an efficient alternative to quizzes or other more elaborate testing approaches (*e.g.* Genet and Sargent 2003; McClintock *et al.* 2010*a*).

Not surprisingly, we found that rare species were correctly identified less often than their common variants among all skill levels (Figure 3.1A and Table 3.4). Interestingly, more skilled observers tended to submit false-positive records of rare species more often than common ones, whereas the reverse was the case for moderately-skilled observers, who incorrectly detected common species more often than rare species (Figure 3.1B and Table 3.5). One explanation for this interaction might be that more experienced observers have a greater familiarity with rarer species than novices and therefore may be aware of a greater number of alternatives for a given vocalization (*e.g.* Faanes and Bystrak 1981).

These results further suggest that naively-modeled data collected mostly from experts may overestimate the occupancy or abundance of rare species. Where similar-sounding rare and common species are not present simultaneously, the nondetection errors associated with these false positives would also underestimate occupancy or abundance of common species. Conversely, surveys using less-skilled volunteers would overestimate common species occupancy and underestimate occupancy for similar-sounding rare species. Hence, our data support existing evidence that heterogeneous mixtures of surveyor skill levels can lead to biased detection and occupancy estimates (*e.g.* Fitzpatrick *et al.* 2009).

In light of these detection biases, survey designers must control for skill level among participants (*e.g.* Kepler and Scott 1981; Genet and Sargent 2003), incorporating rare-species interaction effects as appropriate. This is important when working with both single-visit and single-observer designs (present study), and repeated-sampling designs (the preferred approach; *e.g.* Fitzpatrick *et al.* 2009). Independent of any skill effects, the nontrivial nondetection and false-positive rates we observed also emphasize that neither form of error can be ignored.

Contrary to our expectations, we found no evidence that an incentive to correctly detect rare species contributed to differences in detection error rates across or within skill levels (Tables 3.4 and 3.5). We therefore have no evidence that the intrinsically competitive designs of surveys such as *eBird* (Sullivan *et al.* 2009) – which publishes observers' names alongside their detection records and encourages the detection of rarities – or surveys with informally competitive cultures such as the Audubon Christmas Bird Count (Preston 1958; Butcher *et al.* 1990; Dunn *et al.*

2005; Bonta 2010) – which encourages the detection of large numbers of species – could introduce bias and affect error rates. However, our survey design offered only a weak incentive – in particular, no guarantee of publicity among one's peers – hence, we suggest that these results be regarded as preliminary.

Finally, we found that observers of higher self-assessed skill levels tended to be more confident about the correctness of their identifications than less-skilled observers (Figure 3.2A). These more-skilled observers also tended to have fewer false-positive and more correct responses (Figure 3.1). Thus, the higher confidence of experts was justified in principle. However, our specific measurement of observer confidence was whether observers believed that they had made *zero* detection errors on a given survey scenario, and this specific outcome was actually quite rare. We found a consistent overconfidence among observers of all skill levels (Figure 3.2B). Thus, an apparent increase in the level of observer confidence with increasing self-assessed skill seems to have outpaced the proportionately smaller increase in actual ability, causing the level of overconfidence to remain consistent across observers of different skill levels.

A promising model-based approach to account for the nontrivial instances of both nondetection errors and false positives in detection survey data requires that observers provide a measure of the reliability of each species detection (Miller *et al.* 2011). Because we found widespread levels of overconfidence in our dataset, we believe that a subjective declaration of certainty (*e.g.* a rating of observer confidence from 1 to 10; Larrick *et al.* 2007) for use as such a reliability measure may not be appropriate, and more objective measures such as the anuran chorus-intensity values used by Miller *et al.* (2011) are preferable. For bird surveys, observers could also note the call type (*e.g.* the *chick-burr* call for a Scarlet Tanager [a highly-confident identification] *vs.* its less-distinctive, Robin-like song, a less-confident identification), or more generally, the type of detection method used (*e.g.* heard vs. seen; Miller *et al.* 2011). Recording such detailed detection evidence is not an impractical option, as it has already been successfully implemented on broad scales in several Canadian breeding bird atlases (*e.g.* the Maritimes Breeding Bird Atlas 2006-2010), which require observers to classify detections using a range of breeding evidence codes. Another important complementary strategy is to emphasize to volunteers the value of being conservative with one's species identifications, for instance recording no observations when in doubt (sensu McClintock *et al.* 2010*a*), which can reduce the incidence of false positives arising from overconfidence.

In sum, our results show that an observer's state of mind has important implications for detection errors. Rates of nondetections and false positives vary with species rarity and with observer skill, indicating skill-dependent biases regarding the detection of rare species. Furthermore, overconfidence may be an important factor contributing to these errors. Therefore, approaches to managing these differences that focus on controlling for differences in observer skill and encouraging observer objectivity should improve survey data quality. We hope that this research leads to increasingly fruitful use of the valuable, ongoing contributions of thousands of volunteers.

3.6 Acknowledgments

We thank the volunteer observers who participated in the survey, including members of the Nova Scotia Bird Society and Bird Studies Canada, as well as readers of the e-mail listservers *NatureNS*, *NatureNB*, *PEIBirds*, *Maine Birding*, *Vermont Birds* and *NH-Birds*. We also thank G. Breed and J. Flemming for helpful discussions. Research was conducted under Dalhousie University's Social Sciences and Humanities Human Research Ethics Board approval #2009-2065, and with financial support from NSERC to RGF (postgraduate scholarship) and MLL (Discovery grant).

Chapter 4

Aging Observers and Long-Term Avian Monitoring: Declining Detection Abilities and Population Trajectory Estimates

4.1 Abstract

Long-term wildlife monitoring often involves volunteer observers conducting annual roadside call counts. However, when a volunteer participates for many years, age-related changes to his or her physiology – in particular declines in hearing ability – might add negative biases to apparent population trajectories. Here, we used independent bird survey data from each of the Atlas of the Breeding Birds of Ontario ('OBBA') and the North American Breeding Bird Survey ('BBS') to show systematic declines in detection probabilities (OBBA) and expected counts (BBS) with increasing observer age. In a case study, we then showed how a failure to account for the continuous effects observer age in a model of Golden-crowned Kinglet (*Regulus satrapa*) BBS count data led to more-negative population trajectory estimates. We also tested for the importance of age-related hearing loss, which tends to affect higher frequencies at earlier ages, by asking if our observed detection declines were greater for species with higher call frequencies. We found some evidence of this effect for species with peak vocalization frequencies above 6 kHz. Among these same species, we also found that previously-published, long-term Canadian bird population trend estimates became increasingly negative as vocalization frequencies increased above 6 kHz. Taken together, our results suggest that observer senescence effects are important influences on long-term survey data quality, and that the mechanisms underlying this process include both age-related hearing loss and other physiological changes. Where possible, we recommend that survey designers and modelers account for observer age in future work.

4.2 Introduction

With support from thousands of volunteer observers, a variety of ongoing, long-term ecological monitoring projects have dramatically improved our ability to address important scientific questions (Silvertown 2009; Peters 2010). With data that can span decades and continents (Silvertown 2009), these projects have become important components of wildlife population assessment and management. For instance, data from the annual North American Breeding Bird Survey ('BBS' [1966-present]; Peterjohn 1994) are used as part of avian species-at-risk assessments (Greenberg and Droege 1999; Dunn 2002) and have also helped to characterize the broad-scale effects of introduced species (Cooper *et al.* 2007), diseases (LaDeau *et al.* 2007) and climatic variation (Link and Sauer 2007; Link *et al.* 2008; Wilson *et al.* 2011). Similar projects have also helped to describe the distributions of anurans (Blaustein *et al.* 1994; Lotz and Allen 2007), invertebrates (Kremen *et al.* 2011), and marine life (Goffredo *et al.* 2010), among others. As with any scientific study, however, the quality of the resulting inferences depends on the project's ability to control for errors that mask true biological patterns.

Missed detections, which occur when an animal is present but not detected, are one such class of error affecting monitoring projects (MacKenzie *et al.* 2005; Royle and Link 2006) and can be caused by many environmental (*e.g.* Griffith *et al.* 2010) and observer-specific (*e.g.* Sauer *et al.* 1994; McLaren and Cadman 1999; Alldredge *et al.* 2007*b*, and see Chapter 3) factors. With present statistical techniques, survey designs that involve repeated sampling, and supplementary data collection, modelers can estimate and correct for missed detection rates directly (Royle and Link 2006; Nichols *et al.* 2009; Miller *et al.* 2011, but see Campbell and Francis 2011). Unfortunately, many point count surveys – including most BBS routes – lack this ideal design, and so model-based covariates are often used to account for important sources of variation in missed detection rates (*e.g.* Link and Sauer 2002). With this approach, if all sources of error are recognized as covariates, the 'corrected' counts can then be used as unbiased indices of overall population size (Johnson 2008).

In practise, however, some of these important covariates might be missed. For instance, the current approach to analyzing BBS data taken by the United States Geological Service (*e.g.* Link and Sauer 2002; Sauer and Link 2011) does not explicitly account for changes in detection ability within particular observers over time (except for a first-year 'start-up' effect; Kendall *et al.* 1996; Link and Sauer 2002). Hearing loss is one important physiological reality that changes an observer's ability to detect sounds over time (Gates and Mills 2005), and BBS data consist predominantly of aural detections (Cyr 1981; Faanes and Bystrak 1981). Therefore, if declines in hearing ability as participants age are an important source of missed detection errors, BBS models in their present form could in fact be biased.

Several studies have already argued that changes in observer hearing ability might bias current models of bird species counts (*e.g.* Faanes and Bystrak 1981; Ramsey and Scott 1981; Emlen and DeJong 1992; Simons *et al.* 2007). However, few have tested for such a pattern using field data, and to our knowledge, only Link and Sauer (1998) considered data from multiple observers. In this case, the authors predicted a "43% diminution of counts" for Blue-gray Gnatcatchers (*Polioptila caerulea*) among BBS observers who have conducted surveys for more than 20 years. To our knowledge, this "observer senescence effect" has not been further explored in the published literature. Hence, our understanding of the biases resulting from hearing loss and other such observer senescence phenomena in models of long-term bird survey data is still quite poor.

Age-related hearing loss normally begins after age 30 for both men and women, and includes a broadening range of frequencies beginning above 8 kHz, and progressing to frequencies as low as 1 kHz by age 70 (Figure 4.1; Mayfield 1966; Pearson *et al.* 1995; International Organization for Standardization 2000; Gates and Mills 2005; Van Eyken *et al.* 2007). A second, less-common form of hearing loss involves frequencies in a 'notched' range from 3 to 6 kHz (Nondahl *et al.* 2009; Osei-Lah and Yeoh 2010), which tends to appear more frequently with age (Toppila *et al.* 2001). This 'notched' hearing loss is often associated with noise exposure, and its increased prevalence among older people is likely a function of the cumulative effects of noise and other stressors. Both forms of hearing loss are permanent (Gates *et al.* 2000; Wiley *et al.* 2008; Cruickshanks *et al.* 2010), and they encompass the frequencies produced in many bird vocalizations (Mayfield 1966; Emlen and DeJong 1992), including a group of warblers, nuthatches and flycatchers which we studied here (Table 4.1).

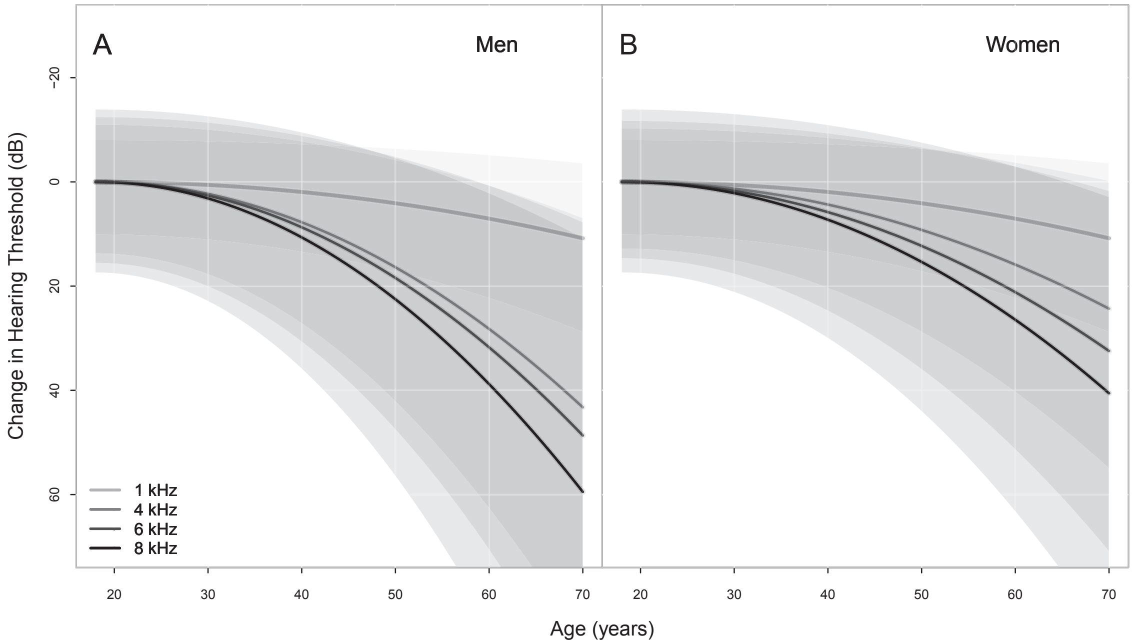

Figure 4.1. International standard for normal changes in hearing thresholds (median values) at standard pure-tone test frequencies (1, 4, 6, and 8 kHz) among (A) men and (B) women of increasing age. Shaded areas are 95% quantiles. Curves are derived from models specified in International Organization for Standardization (2000).

Table 4.1. Table of species used in the various hearing-loss analyses. Standard abbreviations are taken from Klimkiewicz and Robbins (1978). Data source is indicated by an asterisk in the corresponding column. Vocalization frequency information, including peak vocalization frequency (Hz) and power spectrum standard deviation ('SD', as an index of call heterogeneity) are also provided. Frequency range and heterogeneity classifications are also provided, where *low* frequencies are less than 3 kHz, *'notch'* frequencies (corresponding to the audiometric notch related to noise-induced hearing loss) are between 3 kHz and less than 6 kHz, *medium* frequencies are between 6 and less than 7 kHz, and *high* frequency calls exceed 7 kHz. Heterogeneous vocalizations are in the upper 50% quantile of standard deviation values for a broader set that includes 19 additional, unmodeled species (not shown).

Species	Abbrev.	OBBA	BBS (raw)	USGS (BBS) trend	CWS (BBS) trend	Peak Frequency (Hz)	SD	Class
Red-breasted Nuthatch	RBNU	*	*	*	*	2670	514.22	Low Monotone
White-breasted Nuthatch	WBNU	*	*	*	*	2756	329.45	Low Monotone
Brown-crested Flycatcher	BCFL		*			2412	717.27	Low Heterogeneous
Great Crested Flycatcher	GCFL	*	*		*	2584	821.43	Low Heterogeneous
Alder Flycatcher	ALFL	*	*		*	4307	646.83	Notch Monotone
Ash-throated Flycatcher	ATFL		*	*		3101	319.57	Notch Monotone
Black-throated Gray Warbler	BTGW				*	5082	618.78	Notch Monotone
Brown-headed Nuthatch	BHNU		*	*		4393	576.00	Notch Monotone
Cassin's Kingbird	CAKI			*		3273	606.45	Notch Monotone
Common Yellowthroat	COYE	*	*	*	*	4565	593.58	Notch Monotone

Continued on next page

Table 4.1, continued

Species	Abbrev.	OBBA	BBS (raw)	USGS (BBS) trend	CWS (BBS) trend	Peak Frequency (Hz)	SD	Class
Eastern Phoebe	EAPH	*	*	*	*	4823	432.01	Notch Monotone
Eastern Wood-Pewee	EAWP	*	*	*	*	4048	475.55	Notch Monotone
Grace's Warbler	GRWA			*		3876	489.64	Notch Monotone
Gray Flycatcher	GRFL			*		3790	678.03	Notch Monotone
Kentucky Warbler	KEWA		*	*		4910	684.07	Notch Monotone
Lucy's Warbler	LUWA		*	*		5512	630.58	Notch Monotone
Olive-sided Flycatcher	OSFL	*	*	*	*	3273	523.34	Notch Monotone
Orange-crowned Warbler	OCWA	*	*	*	*	5082	562.35	Notch Monotone
Pine Warbler	PIWA	*	*	*	*	4221	532.83	Notch Monotone
Pygmy Nuthatch	PYNU		*	*	*	3790	317.78	Notch Monotone
Say's Phoebe	SAPH		*	*	*	3531	453.42	Notch Monotone
Scissor-tailed Flycatcher	STFL		*			3704	623.82	Notch Monotone
Swainson's Warbler	SWWA		*	*		4996	660.03	Notch Monotone
Vermilion Flycatcher	VEFL		*	*		4048	620.00	Notch Monotone
Western Wood-Pewee	WWPE		*	*	*	3445	405.60	Notch Monotone
Yellow-bellied Flycatcher	YBFL	*	*		*	4134	646.14	Notch Monotone
Acadian Flycatcher	ACFL		*	*		4823	755.32	Notch Heterogeneous

Continued on next page

Table 4.1, continued

Species	Abbrev.	OBBA	BBS (raw)	USGS (BBS) trend	CWS (BBS) trend	Peak Frequency (Hz)	SD	Class
American Redstart	AMRE	*	*	*	*	5943	915.55	Notch Heterogeneous
Black-throated Blue Warbler	BTBW	*	*		*	4221	713.75	Notch Heterogeneous
Black-throated Green Warbler	BTGW	*			*	4393	905.87	Notch Heterogeneous
Black Phoebe	BLPH		*	*		5082	951.26	Notch Heterogeneous
Canada Warbler	CAWA	*	*	*	*	5857	830.75	Notch Heterogeneous
Cerulean Warbler	CRWA	*	*	*		4221	879.96	Notch Heterogeneous
Chestnut-sided Warbler	CSWA	*	*	*	*	5340	1101.86	Notch Heterogeneous
Connecticut Warbler	COWA			*	*	4996	1100.18	Notch Heterogeneous
Dusky Flycatcher	DUFL		*	*	*	4910	763.74	Notch Heterogeneous
Hammond's Flycatcher	HAFL		*	*	*	5857	973.81	Notch Heterogeneous
Hermit Warbler	HEWA		*	*		5168	919.93	Notch Heterogeneous
Hooded Warbler	HOWA	*	*	*		4048	761.85	Notch Heterogeneous
Louisiana Waterthrush	LOWA		*	*		4565	796.63	Notch Heterogeneous
MacGillivray's Warbler	MGWA		*	*	*	4996	708.54	Notch Heterogeneous
Magnolia Warbler	MAWA	*	*	*	*	4910	1283.08	Notch Heterogeneous
Mourning Warbler	MOWA	*	*	*	*	3962	891.97	Notch Heterogeneous
Northern Waterthrush	NOWA	*	*	*	*	4565	1111.92	Notch Heterogeneous

Continued on next page

Table 4.1, continued

Species	Abbrev.	OBBA	BBS (raw)	USGS (BBS) trend	CWS (BBS) trend	Peak Frequency (Hz)	SD	Class
Palm Warbler	PAWA	*	*	*	*	3618	840.98	Notch Heterogeneous
Prairie Warbler	PRWA		*	*		5340	855.07	Notch Heterogeneous
Ruby-crowned Kinglet	RCKI	*	*	*	*	3790	1144.54	Notch Heterogeneous
Townsend's Warbler	TOWA		*	*	*	5512	860.07	Notch Heterogeneous
Virginia's Warbler	VJWA		*	*		5082	696.80	Notch Heterogeneous
Western Kingbird	WEKI		*	*	*	3618	1004.12	Notch Heterogeneous
Willow Flycatcher	WIFL	*	*		*	3618	693.10	Notch Heterogeneous
Wilson's Warbler	WIWA	*		*	*	5771	1122.54	Notch Heterogeneous
Yellow-breasted Chat	YBCH		*	*	*	3876	1212.40	Notch Heterogeneous
Yellow-rumped Warbler	YRWA	*		*	*	4307	832.30	Notch Heterogeneous
Yellow-throated Warbler	YTWA		*	*		5857	732.75	Notch Heterogeneous
Yellow Warbler	YEWA	*	*	*	*	5340	889.48	Notch Heterogeneous
Black-and-white Warbler	BAWW	*	*	*	*	6718	663.48	Medium Monotone
Blue-winged Warbler	BWWA	*	*	*	*	6632	591.23	Medium Monotone
Brown Creeper	BRCR	*	*	*	*	6977	679.02	Medium Monotone
Cedar Waxwing	CEWA	*	*	*	*	6891	314.65	Medium Monotone
Golden-winged Warbler	GWWA	*	*	*	*	6029	424.40	Medium Monotone

Continued on next page

Table 4.1, continued

Species	Abbrev.	OBBA	BBS (raw)	USGS (BBS) trend	CWS (BBS) trend	Peak Frequency (Hz)	SD	Class
Worm-eating Warbler	WEWA		*	*		6546	514.35	Medium Monotone
Cordilleran Flycatcher	COFL		*			6460	973.74	Medium Heterogeneous
Eastern Kingbird	EAKI	*	*	*	*	6202	1171.83	Medium Heterogeneous
Least Flycatcher	LEFL	*	*	*	*	6718	1276.71	Medium Heterogeneous
Nashville Warbler	NAWA	*	*	*	*	6202	961.56	Medium Heterogeneous
Northern Parula	NOPA	*	*	*	*	6891	786.68	Medium Heterogeneous
Ovenbird	OVEN	*	*	*	*	6202	991.24	Medium Heterogeneous
Bay-breasted Warbler	BBWA	*	*	*	*	7321	490.08	High Monotone
Blackpoll Warbler	BPWA	*		*	*	8269	257.59	High Monotone
Cape May Warbler	CMWA	*	*	*	*	7580	375.51	High Monotone
Golden-crowned Kinglet	GCKI	*	*	*	*	7235	680.00	High Monotone
Blackburnian Warbler	BLWA	*		*	*	7666	828.64	High Heterogeneous
Prothonotary Warbler	POWA		*	*		7494	1213.31	High Heterogeneous
Tennessee Warbler	TEWA	*	*	*	*	8958	1216.47	High Heterogeneous

In the BBS, which is a large and influential source of North American bird population data (Sauer and Link 2011), the majority of observers are over 45 years of age (Figure 4.9; and see Wiedner and Kerlinger 1990; La Rouche 2001; Downes 2004; Carver 2009), and they tend to survey routes repeatedly for many years. There has also been an increase in the average number of years served on the BBS since data collection began in 1966, which may indicate that average observer age has also increased (Figure 4.2). Given the age-associated changes in human hearing, we hypothesize that this modern, aging population of BBS volunteers may be detecting fewer birds because of an inability to hear them, independent of real ecological patterns. If so, observed population declines derived from these data may in fact reflect reduced detection abilities, independent of any real biological change. With thousands of volunteers playing important roles in population monitoring in the BBS and similar studies (*e.g.* the Audubon Christmas Bird Count; La Sorte and McKinney 2007), this is a potentially serious issue.

Our goals for this study were to test for the existence and consequences of age-related declines in the detection abilities of long-term bird survey observers, with a focus on hearing loss as a potential mechanism. We used data from two independent volunteer bird survey datasets: the Atlas of the Breeding Birds of Ontario ('OBBA'; Bird Studies Canada *et al.* 2008) and the North American Breeding Bird Survey (Peterjohn 1994) to establish how bird detection probabilities and expected species counts, respectively, tend to change as observers age. We considered the role of hearing loss by simultaneously testing for patterns between age-related changes in detection ability and species vocalization frequencies. We expected to see the strongest effects for species with vocalization frequencies corresponding to common age-associated hearing impairments.

We then asked how age-related changes in observer ability might bias long-term estimates of population change. We first conducted a case study using BBS counts of the Golden-crowned Kinglet ('GCKI'; *Regulus satrapa*), a small songbird chosen because of its high-frequency vocalization (Table 4.1) and relatively high mean abundance. Here, we used data from this species to illustrate how using uncorrected data from aging observers might bias population trajectory estimates. We then

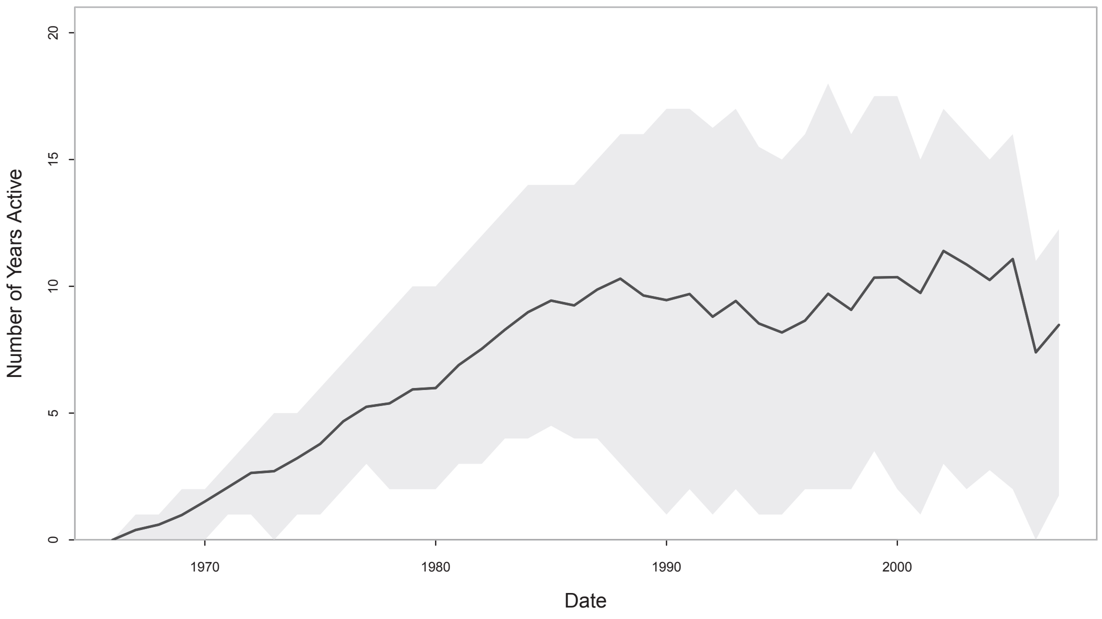

Figure 4.2. Change in mean minimum observer age of Canadian participants in the North American Breeding Bird Survey. 'Minimum observer age' is a relative measure based upon the number of years an observer has participated in the survey ('Number of Years Active'). Shaded area denotes the second and third quartiles of the data.

considered whether hearing loss in particular might be adding error to existing long-term population trend estimates. In this case, we tested for a relationship between previously-published, long-term population trends and the vocalization frequencies of each corresponding species. Similar to our analyses of detection probabilities, we expected that population trend estimates would be more negative in species with vocalization frequencies corresponding to common forms of hearing loss.

4.3 Methods

4.3.1 Calculating and Classifying Vocalization Frequencies

In all analyses, we focused on a group of North American songbirds (warblers, nuthatches, flycatchers) for which we could obtain high-quality vocalization data (Table 4.1). These species form a major proportion of North American breeding bird species (*e.g.* http://www.mbr-pwrc.usgs.gov/bbs/specl09.html), they represent a broad range of vocalization frequencies (Brand 1938), and they are frequently of conservation interest (*e.g.* Faaborg *et al.* 2010).

To test for the influence of hearing loss phenomena (*e.g.* age-related [high frequency] and noise-associated ['notch'] hearing losses) in each analysis, we first determined the peak (*i.e.* dominant) frequencies for the vocalizations of each species following Emlen and DeJong (1992). For each species, we obtained an audio recording of its typical vocalizations (calls and songs) from the Macaulay Library at the Cornell Laboratory of Ornithology (http://macaulaylibrary.org accessed on 08 March 2011) and generated power spectra from each recording using the free software Audacity (Beta 1.3; http://audacity.sourceforge.net/ accessed 08 March 2011). Power spectra display the total energy expended during an audio sample (dB) for each of a contiguous range of narrow frequency bins (*i.e.* 2.00–2.08 kHz, 2.081–2.160 kHz; Figure 4.10). To reduce the effects of any background noise, we converted the log-scale power (dB) values to a linear equivalent (details in Appendix 1). We then noted the peak frequency, defined as the upper bound of the frequency bin with the highest power (Table 4.1). With this approach, the length of the recordings and the number of vocalizations featured in each recording were unimportant, as the power spectra considered the power and frequencies of all sounds present on each recording collectively.

For one of our analyses, we then used these peak frequency values to assign species into one of four vocalization frequency groups corresponding to known hearing-loss thresholds of 3 kHz, 6 kHz and 7 kHz (International Organization for Standardization 2000). Accordingly, species were considered to have 'low' ($<$ 3 kHz), 'notch' ($\geq$ 3 and $<$ 6 kHz), 'medium' ($\geq$ 6 and $<$ 7 kHz) and 'high' ($\geq$ 7 kHz) vocalizations (Table 4.1). We hypothesized that the detection of 'high' vocalizations ($\geq$ 7 kHz) is most likely to change with age (a result of age-related hearing loss; *e.g.* Figure 4.1; International Organization for Standardization 2000; Gates and Mills 2005); whereas 'notch' vocalizations (3 to 6 kHz) may also show age-associated detection declines (cumulative noise-induced and other idiopathic hearing losses; Nondahl *et al.* 2009; Osei-Lah and Yeoh 2010). 'Medium' and 'low' vocalizations lie between the 'high' and 'notch' categories.

'Peak frequency' is most representative of a particular vocalization if the vocalization broadcasts a very narrow range of frequencies overall (Figure 4.10; Ramsey and Scott 1981). By comparison, this single statistic is much less relevant to vocalizations incorporating several high-energy sounds at many disparate frequencies. To account for this difference and focus on the former type, we thus classified vocalizations as being either 'monotone' or 'heterogeneous' according to the range of frequencies found in each power spectrum (standard deviation of power values; Table 4.1). We defined 'monotone' vocalizations as those vocalizations with power spectra having a standard deviation less than or equal to the median value among a group of 94 species initially considered; all other vocalizations having more variable power spectra were defined as being acoustically 'heterogeneous'. Using these heterogeneity classes, we thus expanded the existing four vocalization groups discussed above into eight (*i.e.* 'High Monotone', 'High Heterogeneous', 'Medium Monotone', 'Medium Heterogeneous', 'Notch Monotone', 'Notch Heterogeneous, 'Low Monotone', 'Low Heterogeneous'; Table 4.1). We expected that any relationships between species detections and vocalization frequencies resulting from frequency-specific hearing loss phenomena would be stronger among monotone species.

This simple classification method did not recognize cases where bird vocalizations featured a wide range of frequencies broadcast over a very short time interval (*e.g.* Least Flycatcher [*Empidonax minimus*]) – and so which appear subjectively monotone to the human ear in spite of their having a heterogeneous power spectrum. However, this error did not risk the inclusion of subjectively heterogeneous species in the monotone groups – a more serious error because we were largely concerned with patterns among monotone species only – and so this error was a conservative one.

4.3.2 Determining Observer-Age-Related Changes in Species Detections

Changes in OBBA Detection Probabilities

To estimate the change in detection probability between older and younger observers (defined below), we used data from 43 species surveyed as part of the OBBA which had at least 100 detection records in total (Table 4.1), and for which we were able to determine peak vocalization frequencies (see above). The OBBA is a volunteer survey that divides the entire land area of the Canadian province of Ontario into a grid of

3,324 10 km × 10 km squares, and during two 5-year periods ('first atlas': 1981–1985; 'second atlas': 2001–2005), one to several volunteers per square conducted area searches for breeding evidence of bird species during the spring and summer months, with a minimum effort of 20 party-hours per square.

Working with atlas squares sampled during at least two separate years by one or more observers, we inferred species detections as occurring when an atlasser reported any evidence of a given species in a given atlas square. Conversely, we inferred nondetection for a given atlasser and species by determining all squares visited by an atlasser, and assigning zeroes ('no detection') to all species that were not reported there (sensu Kéry *et al.* 2010).

We used publicly-available data sources, including OBBA results web pages and field naturalist groups' newsletter reports, to determine the approximate ages (under-40, 40–50, or over-50) during the midpoint of the second atlas (2003) for 350 of 1,230 atlassers (demographic data were not available for most observers). Although our primary interest was in measuring age-related patterns of detection ability, we also recognized that gender could have an influence, because men tend to lose their high-frequency hearing sooner than women (Figure 4.1), and so we recorded gender as well. We also corrected for observer effort, both by excluding records with zero-values and by modeling species detectability with effort as a covariable.

We could not explicitly distinguish between casual, "backyard" observations and more-distant field searches within a given atlas square, the former of which might be more often associated with older, less-mobile birdwatchers. Any such relationship, if widespread in the data, could confound age-related differences in observer sensory abilities (of interest here, and relevant to fixed-protocol surveys like the BBS) with age-related differences in observer mobility. However, we were unlikely to successfully determine an observer's age and hence, include that observer's data, unless he or she had a substantial field naturalist group presence (for instance, sufficient to warrant publishing his or her name and photograph in a newsletter), and in our experience, active field naturalist group participation implies an ability and a preference to visit sites further afield than a backyard. Hence, we suspect that this potential confounding influence is not widespread in the data used here.

To model the effects of observer age on detection probability, we constructed hierarchical occupancy models of the resulting dataset (Royle and Kéry 2007; Royle and Dorazio 2009) in WinBUGS 1.4.3 (Lunn *et al.* 2009) and R 2.13.0 (R Development Core Team 2011) using the R package *arm* (Gelman *et al.* 2010) on a PC running Windows 7. Hierarchical occupancy models simultaneously estimate species presence ('occupancy') and detection probability (conditional on occupancy), along with the effects of specified covariates, from detection/nondetection datasets that have been repeatedly sampled at a set of locations (*i.e.* atlas squares). Consistent with the assumptions of OBBA design, we assumed that occupancy did not change for a given atlas square during each of the 5 years of an atlassing period, and so treated each atlas square as a sampling unit and each sampling year as a within-observer replicate.

In the models, we used observers from two age groups, observers under 40 (18 women, 65 men) and observers over 50 (64 women, 203 men), and expected that the older cohort would have functionally reduced hearing compared to the younger one, on average (Figure 4.1). We avoided using records for atlassers we believed to be aged between 40 and 50 years ($n = 60$) in order to preserve such an expected functional distinction (sensitivity analyses later validated this concern; see below), and to account for any errors in age-determination. Each hierarchical occupancy model consisted of an occupancy component, which predicted true occupancy in the second atlas as a function of detection in the first atlas, and a detection component, which predicted detection (conditional upon occupancy) as a function of effort (survey hours), observer gender, observer age (over 50 *vs.* under 40), and random observer variation. Specific formulations of the occupancy models and Bayesian priors used are discussed in Appendix 2.

In each converged occupancy model, the 'β_2' 'observer age' parameter (see equation 4.5 in Appendix 2) corresponded to the difference in detection probability on the logistic scale between observers younger than 40 and older than 50 for a given species. The mean of all β_2 estimates describes how observer detection ability is expected to change with age among all species considered.

We then described the role of hearing loss in driving age-related changes in this detection probability statistic, if any, by constructing an additive model ('GAM';

Wood 2006) using the R package *mgcv* (Wood 2006) to predict the 43 estimates of β_2 as a function of the peak vocalization frequencies of each corresponding bird species. Compared to parametric approaches, which require polynomial curve orders to be defined *a priori*, GAMs can fit nonparametric smooth functions predicting the most likely nonlinear relationships between x (*i.e.* vocalization frequency) and y (*i.e.* β_2), with optimized amounts of 'wiggliness' (Wood 2006). These smooth functions are intended to be viewed by the modeler; summary statistics alone are inadequate to describe their findings (here, p-values correspond to the probability that a smooth function exists by random chance alone).

We considered relationships between β_2 and vocalization frequency for the monotone and heterogeneous vocalizations separately, and weighted the datapoints according to the inverse of the variance of their posterior distributions (*i.e.* their uncertainty; estimated earlier by WinBUGS). If age-related hearing loss is an important mechanism leading to age-related detection declines, we expected the magnitude of the β_2 values to decline with increasing vocalization frequency, mirroring the hearing threshold curves in Figure 4.1. Declines in the 'notched' region would similarly correspond to an influence of noise-induced hearing loss. We also expected to see more-pronounced patterns for the 'monotone' species groups, because in this case, the peak vocalization frequency more-closely corresponds to the principal frequency heard by the observer.

To justify our exclusion of observers from the 40–50 age group, and furthermore to validate whether observer age influences detection ability, we also tested for the sensitivity of the values of β_2 to the inclusion of observers of borderline age (*i.e.* ages 40–50). We re-fit the occupancy models as described above, except here using data from observers of all ages (while retaining the age-50 cutoff), and then compared matched pairs of these new β_2 estimates to their earlier estimates. If there are important differences in detection ability that develop progressively by age 50, we would expect to see a smaller overall change in detection ability between the under-50 and over-50 cohorts, compared to the changes previously measured between the under-40 and over-50 cohorts.

Expected BBS Counts Derived From BBS Data

Next, to compare the patterns observed above with an independent dataset, we determined how bird counts changed with increasing observer age on the BBS. Like the OBBA, the BBS is a multi-year, omnibus bird survey. Here, it is conducted by skilled volunteers annually during the breeding season. In contrast to the OBBA, most BBS survey stops are not replicated within survey cycles or among multiple observers. The BBS uses a set of permanent, 39.4 km road transects ('routes') which are divided into 50 stops placed at regular ($\sim$ 800 m) intervals. Most BBS routes are sited randomly within North American physiographic subregions ('strata'; *e.g.* 'Sierra Nevada'; 'St. Lawrence River Plain') and within degree blocks of latitude and longitude (Sauer *et al.* 2003), and so have a nested random structure. Survey routes continue to be added to the BBS as a whole; the oldest routes have been monitored annually since 1966.

In the raw BBS count dataset, observers are assigned a unique identification number which persists throughout their years of service. We used these identification numbers to determine a measure of "minimum observer age", defined as the number of years since the first year an observer served on any BBS route (sensu Faanes and Bystrak 1981). Within observers, minimum observer age is correlated with actual observer age – our latent variable-of-interest – by definition. We recognize that this measure is less-precise than true age, however for simplicity, we refer to 'minimum' observer ages, which range from 1 to 39 in the data, as 'observer ages'.

We omitted data from the early years of the BBS, and instead used data collected in Canada and the USA between 1970 and 2007 by single observers under suitable weather conditions. These omissions avoided potential problems with low observer quality in the early years of the survey (*e.g.* Sauer *et al.* 1994), as well as problems with 'anomalous results' with early data from Canadian survey routes (http://ec.gc.ca/reom-mbs/default.asp?lang=En&n=E8974122-1 accessed on 25 March 2011). Observer ages were calculated using the original, complete dataset, which began in 1966.

To account for changes in each species' population abundance occurring alongside changes in observers' detection ability, we needed replicated time series for each location. To achieve this effect with the unreplicated BBS data, we estimated population trajectories at the level of the physiographic stratum, using count data from the individual survey routes as replicates. We required that at least 3 separate observers be associated with a given stratum before that stratum was included in the analysis. Similarly, to ensure that the age and population effects under study were not confounded, we required that for each species and stratum analyzed, observer age and calendar year were correlated by no more than 0.7 (Pearson correlation). We also worked exclusively with contiguous observer-route time series of 10 years or longer, both to minimize biases that could result from any gaps in temporal coverage (*e.g.* Sauer *et al.* 1994), and to capture age-related changes in detection ability.

Because raw BBS data do not include zero-counts for any species, we added relevant zeroes in the same manner as was done with the OBBA dataset (sensu Kéry *et al.* 2010). The presence of zero values on a route-year time series did not affect whether it was considered contiguous or not. To avoid problems with estimating zero-values on the (logarithmic) scale of the linear predictor in these models, we also added a value of 0.5 to all counts (Sauer *et al.* 1996).

Volunteer BBS observers often perform relatively poorly during their first year on a survey route compared to later years; this phenomenon can inflate population trend estimates if the first year of data is included (Kendall *et al.* 1996). To avoid confounding this pattern with hearing loss phenomena, we excluded the first years' datapoints (mean 6.1% of records per species) for each observer-route combination. Final datasets for each of 65 species meeting these requirements (and for which we had vocalization frequency data; Table 4.1) ranged in size from 37 (Lucy's Warbler [*Oreothlypis luciae*]) to 6692 (Common Yellowthroat [*Geothlypis trichas*]) route-years of data, with a median 1077 records (mean 1608 ± 1689 [SD]).

We used overdispersed Poisson generalized additive mixed models ('GAMMs'; Wood 2006) in R package *gamm4* (Wood 2011) to model the nonlinear change in BBS counts with observer age, while controlling for both among-observer effects and continuous changes in species counts with calendar year within physiographic strata (*i.e.* 'population' changes). GAMMs are extensions of GAMs which incorporate additional random-effects structures to the model to account for group-specific deviations from overall means ('random intercepts') and from overall trends ('random slopes').

These models are suitable for the hierarchical structure of BBS data. As with GAMs, the ‘significance’ of a smooth function alone can only establish whether there is a nonzero pattern to the data – ‘significant’ GAMMs are not necessarily unidirectional (*i.e.* exclusively increasing or decreasing). Hence, GAM and GAMM smooths are visual instruments, and their shapes must always be examined in order to obtain a complete result.

We aggregated the species-specific GAMMs produced here by building a second uncertainty-weighted GAMM describing the overall proportional change in modeled BBS counts (relative to each species’ values at age-1) for groups of species of the same vocalization frequency and heterogeneity groups defined earlier. Finer details of this modeling process are outlined in Appendix 3.

4.3.3 Species Detection Probabilities and Long-Term Population Trend Estimates

Golden-crowned Kinglet Case Study

We used BBS counts of the Golden-crowned Kinglet (‘GCKI’) as the subject of a case study to examine how failing to control for observer age could affect estimated population trajectories. Our goal was to model population trajectories in a manner consistent with established techniques (Link and Sauer 2002; Sauer and Link 2011), and then to compare the resulting estimates made both with and without a continuous correction for minimum observer age.

Here, we excluded continuous observer age effects from the GCKI GAMM produced in the earlier BBS count analysis (see above) to produce a model which approximated the hierarchical Bayesian modeling methods presently used by the US Geological Service (‘USGS’; Link and Sauer 2002). The only major difference between this second GAMM approach and the hierarchical Bayesian models is the GAMMs’ use of a smooth function for calendar year in place of parametric terms to describe annual changes. The numerous parametric terms in the USGS approach should serve roughly the same purpose as a continuous GAMM function, and so both our GAMMs (modeled without observer age corrections) and the USGS hierarchical Bayesian models should produce largely equivalent population inferences from the same initial dataset.

Accordingly, we compared the shapes of the resulting smooth functions for 'population' trajectories ($f_2(l)_j$ in Equation 4.8 in Appendix 3) between GAMMs built with and without age corrections for three physiographic strata that represent both apparently stable and apparently declining 'uncorrected' population trajectories (Northern Spruce-Hardwoods [stable], Sierra Nevada [declining], South Pacific Rainforests [declining]). Because we used the BBS data subset discussed earlier, our population estimates do not necessarily reflect true biological patterns (*e.g.* Sauer *et al.* 1996), but rather they illustrate the directional effects of observer age corrections on population trajectory inferences under severe circumstances (*i.e.* where most modeled observers experience long-term aging).

The Influence of Vocalization Frequencies on Population Trend Estimates

Finally, we determined if species vocalization frequencies – and hence, hearing loss – might have influenced broad-scale population trend estimates. If hearing loss is an important predictive factor, we expected to see greater estimated population declines among species with vocalization frequencies associated with hearing loss. Here, we considered Canada-wide population trend statistics produced by both the USGS (http://www.mbr-pwrc.usgs.gov/cgi-bin/atlasa09.pl?CAN&2&09 accessed on 08 March 2011), and by the Canadian Wildlife Service ('CWS '; http://www.cws-scf.ec.gc.ca/mgbc/trends/index.cfm?lang=e&go=info.SpeciesListByProvince&provid=0 accessed on 08 March 2011). Both sets of trends are calculated by their respective agencies using area-weighted, Poisson-modeled BBS count data, where estimated 'trend' values correspond to the estimated exponential rate of change of a population from the beginning to the end of the survey period modeled. However, fine details of these strategies are not equivalent. Thomas and Martin (1996) showed that agency-specific differences in such analysis strategies (*i.e.* different geographic weighting schemes) can lead to important differences in trend magnitude and significance. Current trend estimation strategies have improved since 1996 among both agencies, but remain divergent for other reasons (C. Francis, pers. comm.). Here, we wanted to determine if there was an agency-independent (*i.e.* common) effect of vocalization frequency among each set of trends, and so considered both sets.

We built single-parameter GAMs relating each of USGS and CWS population trends with species vocalization frequency, specifying separate thin-plate regression spline smoothers for monotone and for heterogeneous vocalizations. The USGS dataset supplied 95% credible intervals about the trend estimates; consequently, we treated the width of these intervals as a measure of error and weighted datapoints according to their inverses. Similarly, we used the supplied number of BBS routes incorporated into each CWS trend prediction as a corresponding weight in the CWS GAM. We used population trend estimates that spanned the longest available timespan in each case, which was from 1966 to 2009 for the USGS trends ($n = 50$ warbler, flycatcher and nuthatch species for which we had vocalization frequency data), and from between 1970 and 1973 to 2009 for the CWS trends ($n = 52$ species). We excluded one CWS trend (Bohemian Waxwing [*Bombycilla garrulus*]) which was valid only for 1986 to 2009.

4.4 Results

4.4.1 Determining Observer-Age-Related Changes in Species Detections

Changes in OBBA Detection Probabilities

The average of the β_2 estimates among all 43 species was negative (mean -0.66 ± 0.81 [SD] on the logit scale; median -0.48; Figure 4.3). These values were not normally-distributed, and only 4 of the 43 β_2 values were greater than zero. Hence, the broad standard deviation is driven by a negative skew to the data, which are almost-entirely below zero. When modeled in the GAM, the intercept term was significantly negative ($p < 0.001$), which indicated a significant overall decline in detection ability among older observers, on average. On a species-specific basis, thirteen of the 43 species considered (30%; BAWW, BBWA, BTGW, COYE, CSWA, GCKI, NAWA, OSFL, OVEN, RCKI, WIWA, YBFL, YRWA; see Table 4.1 for full names) showed 'significant' declines in detectability between younger and older OBBA observers (*i.e.* 95% Bayesian credible intervals of β_2 coefficients did not contain zero). Gender had a less-important influence on detection probability; in this case, the mean effect was much closer to zero (mean 0.17 ± 0.59 [SD]), and seven of the 43 (16%) species showed 'significant' effects of being male on detectability. Contrary to our physiological expectations, most of the gender coefficients (effect of being male) were greater than zero (34 of 43).

Among each of the monotone and heterogeneous species groups, and considering all peak vocalization frequencies, age-related detectability change was not significantly related to the vocalization frequencies (GAM 'slope' smooth terms: $p = 0.297$ [monotone], $p = 0.597$ [heterogeneous]). However, the shape of the curve for monotone species suggested declines in detectability between 3 kHz and 6 kHz ('notch' frequencies), and beyond a threshold of approximately 6 kHz ('medium' and 'high' frequencies; Figure 4.3A). To test for age-related changes in detectability in the higher ($\geq$ 6 kHz) frequency range exclusively, we built a *post hoc*, similarly-weighted linear model predicting age-related changes in detectability as a function of peak vocalization frequency, using only those species with monotone vocalizations above 6 kHz. This model showed a significant linear decline ($p = 0.034$, $n = 9$).

The sensitivity analysis tested how including observers aged 40–50 in the 'younger' age group affected the relative change in detection ability between 'younger' observers and those over 50 ('older' observers). Results showed that when observers between ages 40 and 50 were included in the models as 'younger' participants, the the differences in detection ability between the 'younger' and the 'older' groups (*i.e.* the β_2 values) tended to diminish in magnitude (Figure 4.4). This points to a robust effect of observer age on the detection probability of bird species.

Expected Counts Derived from BBS Data

Model-estimated BBS counts declined significantly (GAMM smooth term $p < 0.05$) over 39 years of increasing observer age for all vocalization frequency groups except Low Monotone ($p = 0.111$) and High Heterogeneous ($p = 0.085$) species (Figure 4.5). Among the significant declines estimated, the greatest changes were among the low-frequency, heterogeneous species (BCFL, GCFL, Table 4.1), which decreased by 66.5%, and the medium-frequency, heterogeneous species (EAKI, LEFL, NAWA, NOPA, OVEN; Table 4.1), which decreased by 59.2% over the 39 years sampled. The smallest significant changes in counts were declines of 34.1% among high-frequency monotone birds (BBWA, CMWA, GCKI, Table 4.1), and 34.3% among notch-frequency monotone birds ($n = 18$ species; Table 4.1) over that same age range (Figure 4.5).

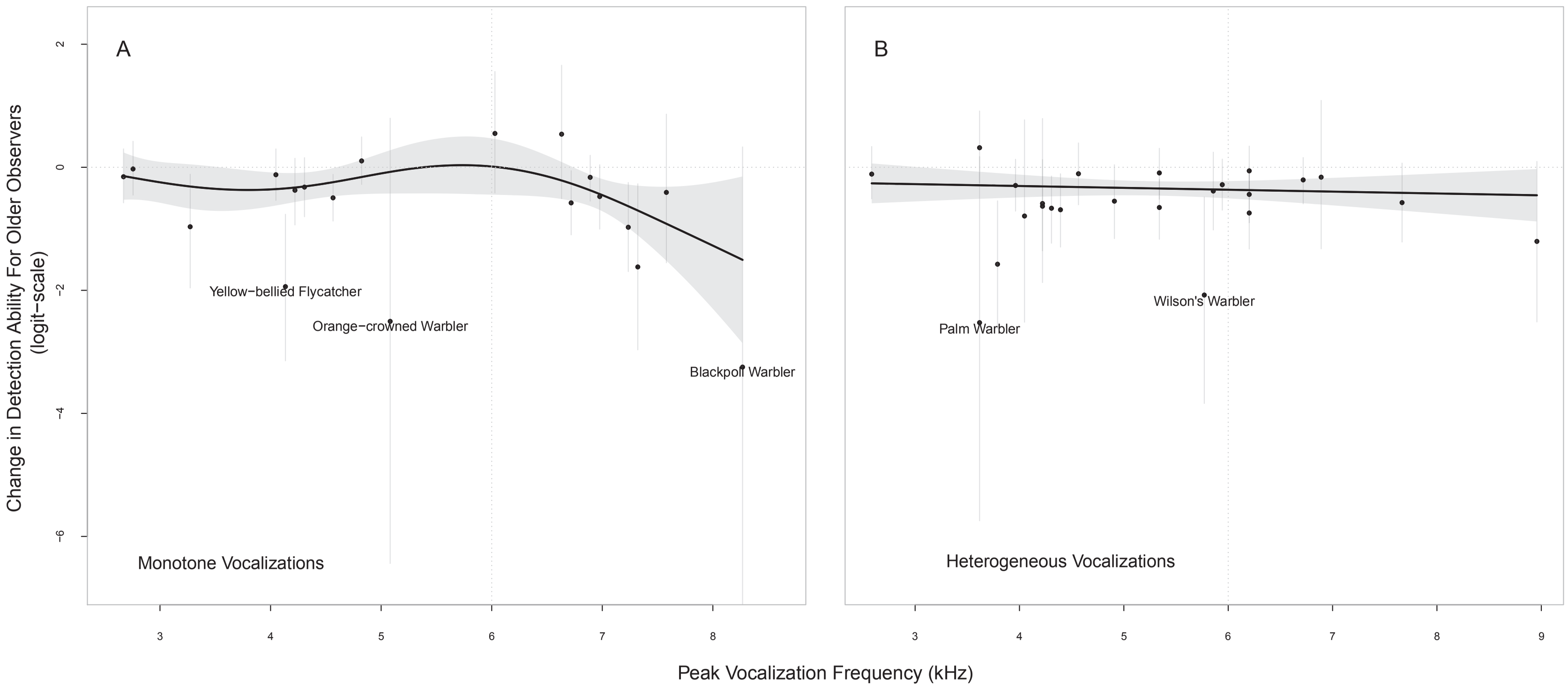

Figure 4.3. Logit-scale difference in species detection probability between an observer over age 50 and an observer under age 40, determined from species-specific hierarchical occupancy models of data from the Atlas of the Breeding Birds of Ontario, as a function of each species' peak vocalization frequency, and grouped by vocalization variability ('Monotone' [A] and 'Heterogeneous' [B]). Smoothed curves are GAM fits, weighted by the inverse of the variance of each datapoint (uncertainty displayed as 95% credible interval lines here), plus the model intercept. Shaded areas are 95% pointwise confidence bands about the smooth term and the model intercept. Dotted reference lines are plotted at $y = 0$ (no difference in detection between younger and older observers) and at 6 kHz (the threshold for "medium" frequencies, as defined in the text).

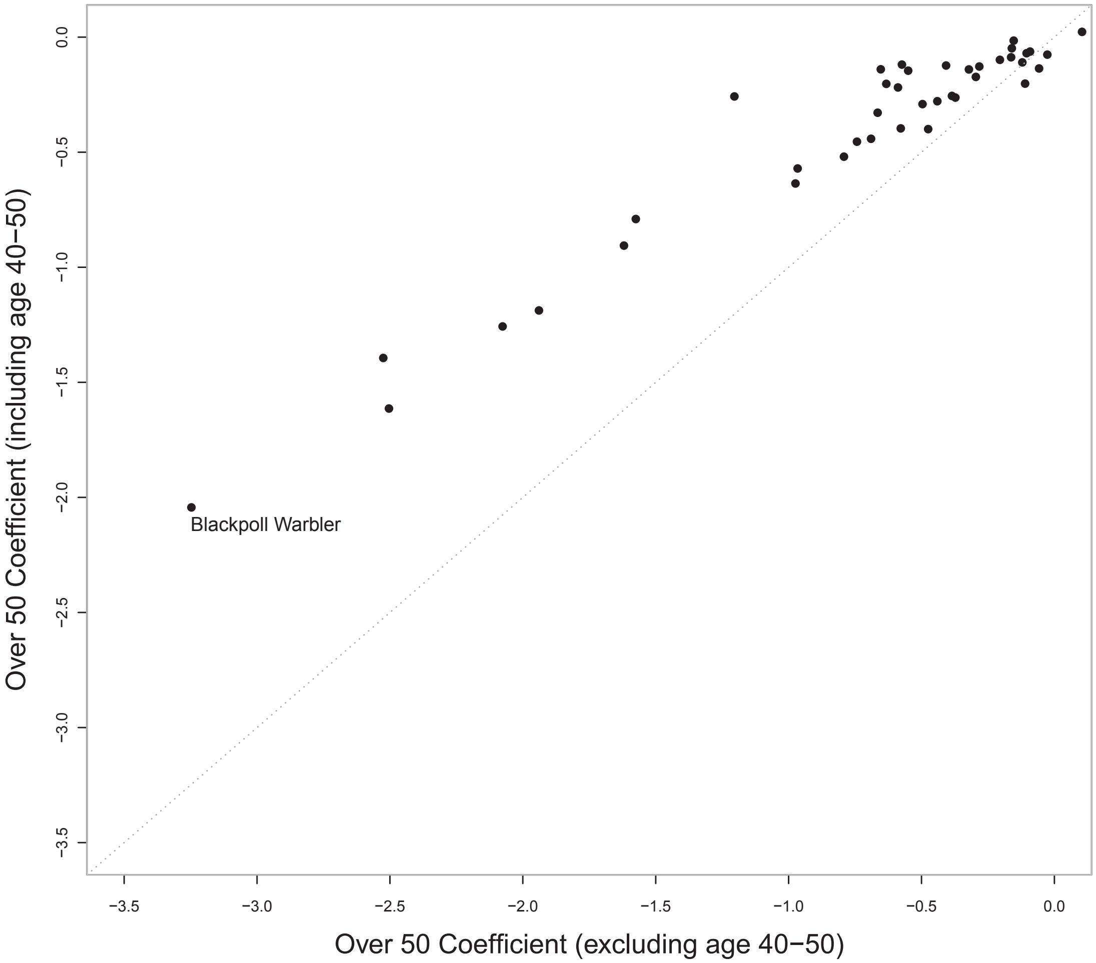

Figure 4.4. Sensitivity of 'β_2' age-related detectability change coefficients (hierarchical occupancy models) to the age structure of the old and young age groups in the modeled data. Coefficients were generated from identical occupancy models using data that either lacked the 40–50 age group as part of its 'young' category (principal modeling approach; x-axis) or included these observers (y-axis). The reference line (dotted) has a slope of 1. When the additional group of middle-aged observers are included in the 'under-50' category (y-axis), the age-related detectability differences are pushed closer to zero.

The increasing uncertainty at the upper range of observer ages (Figure 4.5) reflects the smaller sample sizes in this area.

Golden-crowned Kinglet Case Study

Model-estimated BBS counts of the Golden-crowned Kinglet were expected to decline by nearly 7 birds per observer-route time series after 30 years of aging, all else being equal (Figure 4.6D). Without a correction for such an effect, population trends as inferred visually from the smooth functions for calendar date appeared stable in the Northern Spruce-Hardwoods stratum (Figure 4.6E), and declining in the South Pacific Rainforests and Sierra Nevada strata (Figure 4.6G and I). After correcting for observer age (essentially a vector subtraction of Figure 4.6D from each population smoother), inferred population trajectories became more positive. Specifically, the 'corrected' Northern Spruce-Hardwoods stratum now showed a significant population increase, the South Pacific Rainforests stratum now appeared stable, and the apparent Sierra Nevada decline was less steep.

4.4.2 Species Detection Probabilities and Long-Term Population Trend Estimates

The Influence of Vocalization Frequencies on Population Trend Estimates

There were significant relationships between monotone vocalization frequencies and long-term, Canada-wide population trends for each of the USGS (GAM $p = 0.048$; Figure 4.7A), and CWS datasets ($p = 0.008$; Figure 4.7C), where population trends declined among species with increasing 'medium' and 'high' ($\geq$ 6 kHz) peak vocalization frequencies, and at the midpoint of the 'notched' range (3 to 6 kHz). By contrast, there were no significant relationships between heterogeneous vocalization frequencies and population trends (USGS $p = 0.928$; CWS $p = 0.568$; Figure 4.7B and D). As a whole, the monotone and heterogeneous patterns were also visually similar to those we observed between detection probabilities and vocalization frequencies (*i.e.* Figure 4.3), and they indicated a tendency for species having each of notched and (especially) higher ($\geq$ 6 kHz) monotone frequencies to have more negative long-term population trends. Ignoring uncertainty in the estimated detection probability

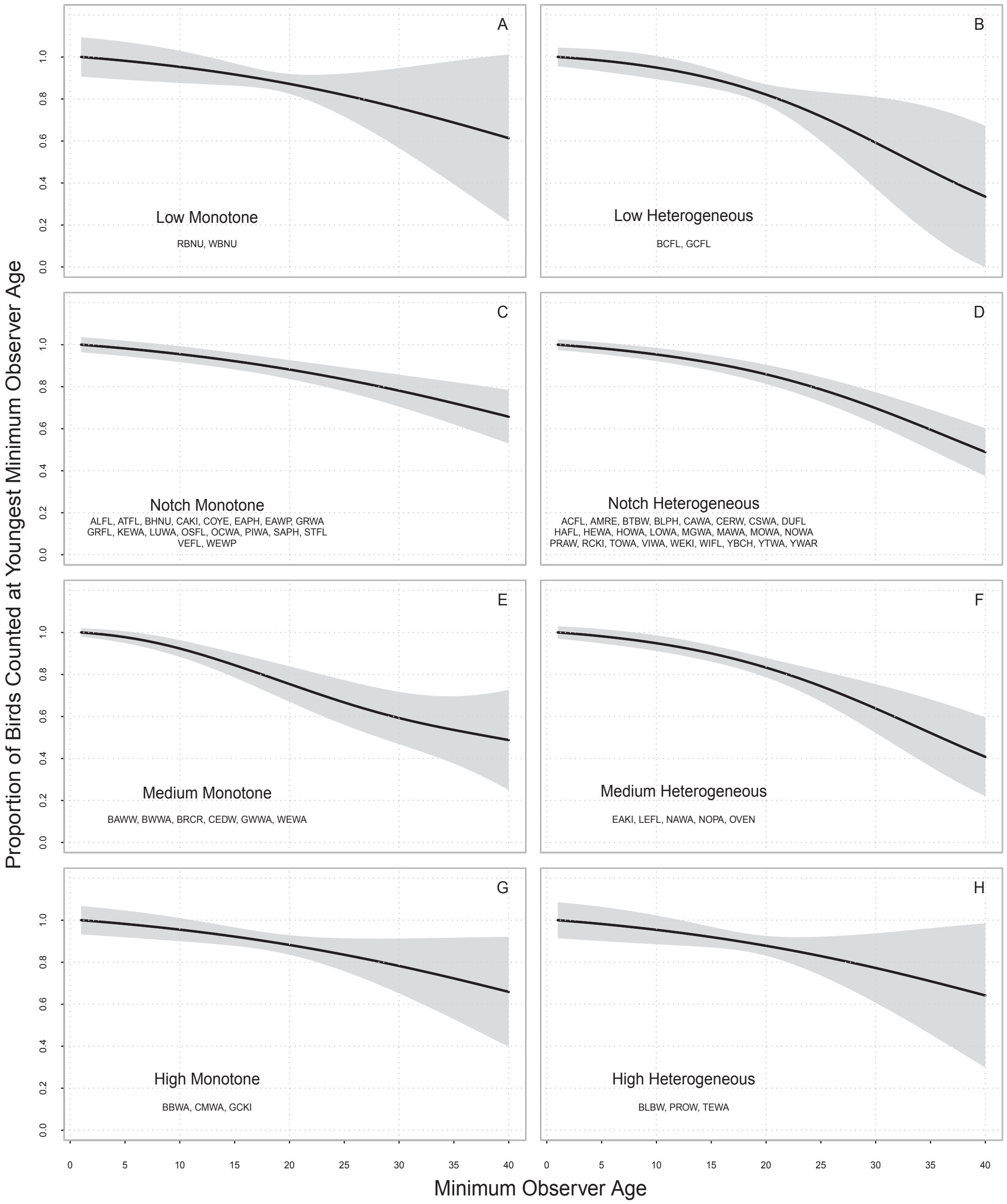

Figure 4.5. Estimated proportional changes in BBS counts with increasing minimum observer age, relative to the count at time zero, grouped by species vocalization groups. Vocalization groups reflect the peak frequency of a typical set of vocalizations for that species, and whether these vocalizations tend to feature a single sound ("Monotone") or a highly-variable set of sounds ("Heterogeneous"). Standard abbreviations for species names (Klimkiewicz and Robbins 1978) belonging to each vocalization group are listed on each panel (also see Table 4.1). Shaded areas are 95% pointwise confidence bands.

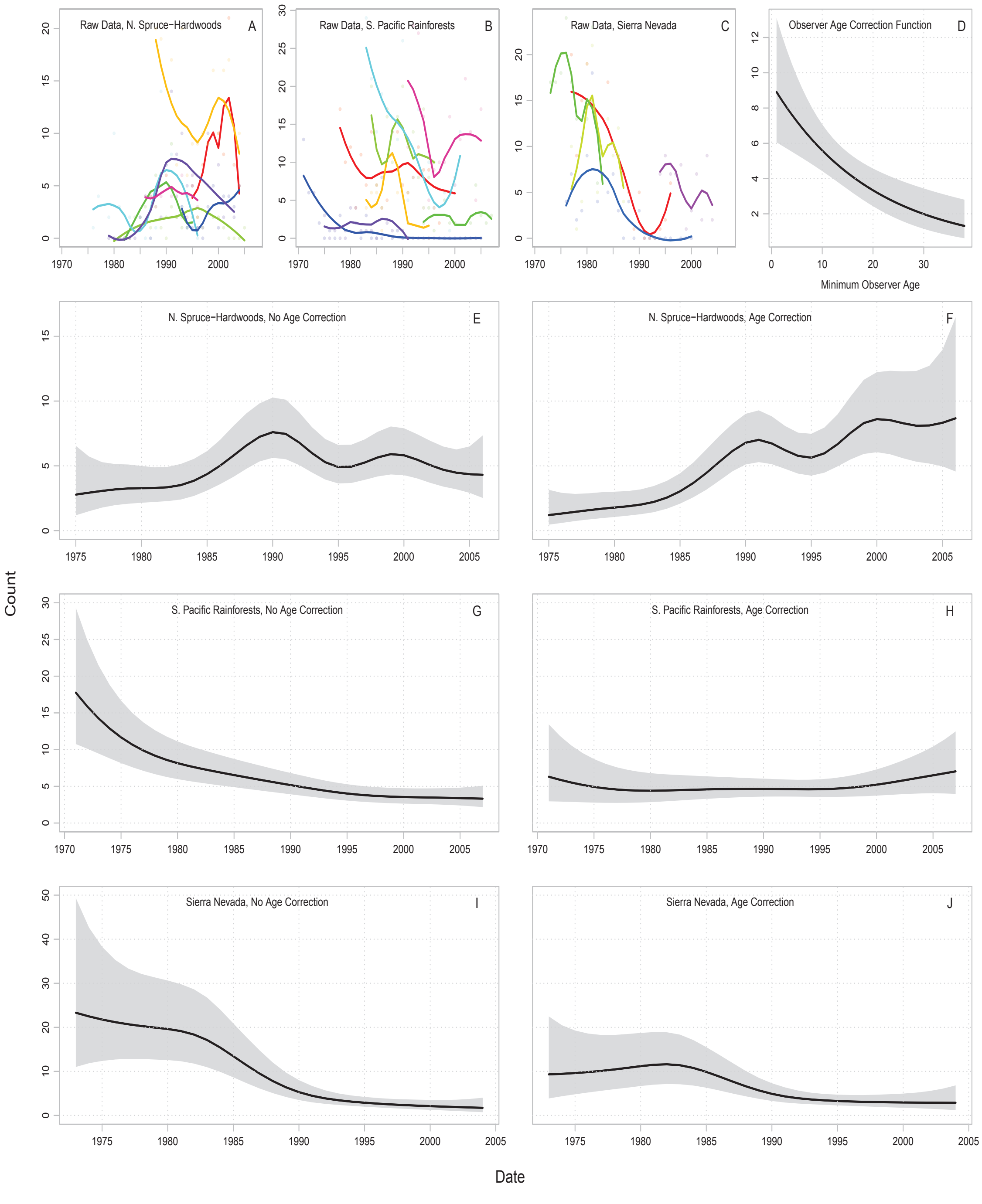

Figure 4.6. Case study of Golden-crowned Kinglet (*Regulus satrapa*) BBS counts, modeled as GAMMs with and without corrections for observer age. Panels A through C show raw data (with loess curves) for a random subset of observer-within-route time series within each of 3 selected physiographic strata. Panels E, G and I show modeled ‘population’ (calendar year) trends, correcting for first-year and among-observer/route effects. Panels F, H and J show ‘population’ trends after making a correction for observer age (shown explicitly in Panel D) in place of the first-year correction. Shaded areas are 95% pointwise confidence bands.

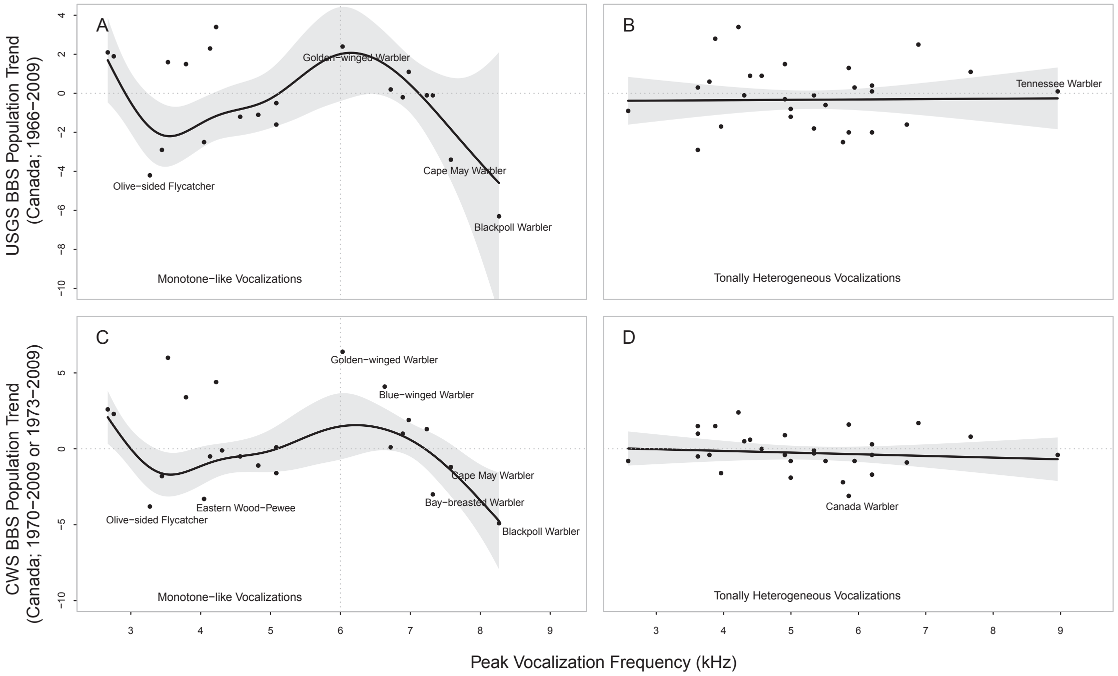

Figure 4.7. Additive model of Canada-wide population trends (1966 to 2009; calculated by the United States Geological Service; panels A and B), and by the Canadian Wildlife Service (1970 to 2009; panels C and D), as a function of each species' peak vocalization frequency (pitch), modeled separately for species with largely single-frequency vocalizations ('Monotone'; panels A and C) and for species with highly-variable vocalization frequencies ('Heterogeneous'; panels B and D). Shaded areas are 95% pointwise confidence bands about the model smooth term plus the model intercept.

change variable, we measured this latter relationship explicitly using Pearson correlations, and found significant patterns for both the USGS data ($r = 0.79$, $p = 0.012$; $n = 9$; Figure 4.8A) and the CWS data ($r = 0.89$, $p = 0.001$; $n = 9$; Figure 4.8B).

4.5 Discussion

Using data from both detection-nondetection (OBBA) and point-count (BBS) surveys; we found several lines of evidence for age-related declines in bird detection abilities among volunteer observers. On average, OBBA observers over age 50 had lower detection probabilities compared to observers under age 40, there were more pronounced detection probability declines between more distant OBBA age groups (*i.e.* Figure 4.4), and there were near-universal declines in expected BBS counts with

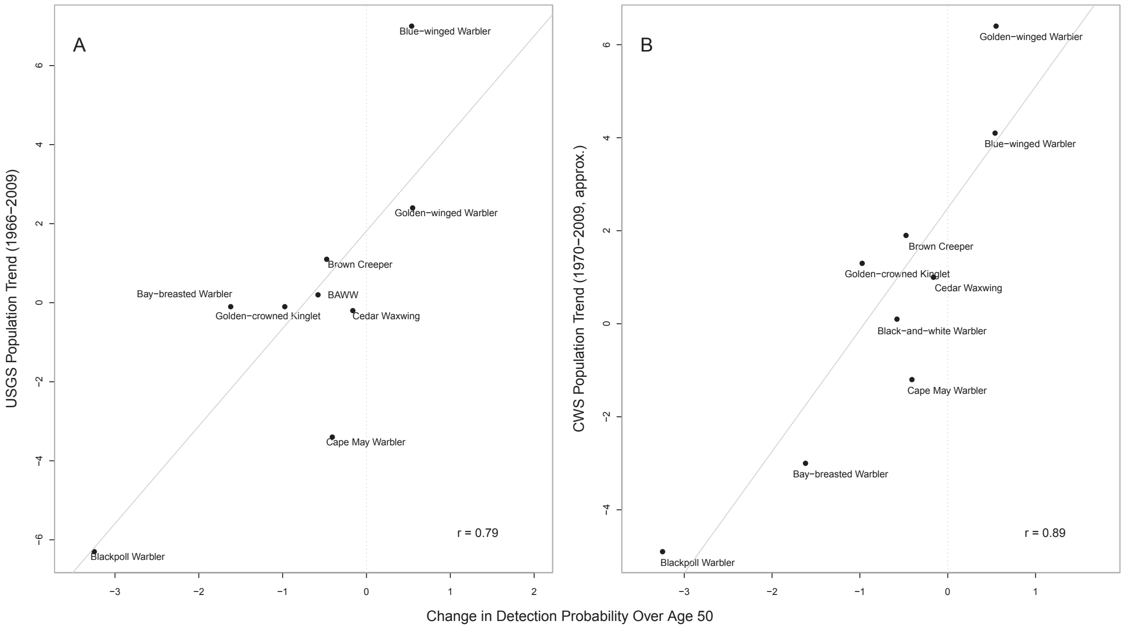

Figure 4.8. Detection probabilities estimated by the occupancy models plotted against (A) USGS and (B) CWS population trends for Canada, for those species common to the two analyses having monotone vocalizations of medium ($\geq$ 6 kHz) or high frequencies. Solid lines correspond to linear regression fits which ignore uncertainty in the detection probability (x-axis) values. Both regression slopes are significantly ($p < 0.05$) different from zero. 'BAWW' is the Black-and-white Warbler (*Mniotilta varia*). This figure combines information from Figures 4.3 and 4.7.

increasing observer age. Among monotone species with peak vocalization frequencies exceeding 6 kHz, age-related detectability changes in the OBBA showed a significant linear decline as peak frequency increased; detection declines may also be occurring at 'notched' frequencies of 3 to 6 kHz. Collectively, these data suggest that observer senescence is an important factor affecting the quality of data from volunteer birdwatchers, and that common patterns of hearing loss play a role in this overall process . However, the concurrent declines in detection ability we observed at other frequencies (*e.g.* Figure 4.5) suggest that other mechanisms are also involved.

Using real data, we found that failing to account for changes in observer detection ability can underestimate population increases, and perhaps more importantly, overestimate population declines. We lastly found indirect evidence of such a bias in previously-published population trend estimates, which tended to be lower as monotone vocalization frequencies above 6 kHz increased, and towards the midpoint of the 'notched' frequency range (3 to 6 kHz), suggesting an uncontrolled confounding effect of age-related hearing losses.

The estimated declines in BBS counts we observed were consistent with data from similar, previous research (Link and Sauer 1998). Here, we showed declines ranging from 34% to 67% of the original counts of more than 60 species considered collectively over 39 years (Figure 4.5), whereas Link and Sauer (1998) estimated a 43% decline in Blue-Gray Gnatcatcher counts among observers after 20 years (our dataset did not include the Gnatcatcher). Using our approach to classifying vocalizations by peak frequency, this gnatcatcher would fall into the Notch Heterogeneous category, for which we estimated a 14.1% decline in counts over 20 years. The smaller value predicted here may result from the large number of species incorporated into this calculation ($n = 25$). In our opinion, either value is large enough to be concerning.

Because the declines in species detections with observer age in both the OBBA and the BBS datasets occurred across most frequencies and in both heterogeneity groups – for instance, the greatest modeled declines in BBS counts were for the Low and Medium Heterogeneous species – we believe that multiple senescence effects, including high-frequency hearing losses, are at work, with hearing-loss effects being most important for species with monotone vocalizations. Normal aging can involve impairments in memory, cognitive speed and vision (Morris and McManus 1991).

Alongside hearing impairments, these factors might each contribute to greater numbers of missed detections independent of bird vocalization frequencies, for instance by limiting one's abilities to (i) simultaneously detect and transcribe species calls, (ii) to recognize multiple, overlapping species calls, and (iii) to identify non-vocal, cryptic species by eye. Hence, while higher-frequency and notched monotone species might be most prone to the effects of age-related hearing loss, all species are probably vulnerable to some form of age-related detection decline. Similarly, because behaviour and visual cues also play a role in bird detection, and can vary from species to species (*e.g.* variable species 'conspicuousness'; Stewart 1954), some species with easily-audible vocalizations might be detected less-often than easily-heard, but cryptic species who sing infrequently (Alldredge *et al.* 2007*a*), and this could explain some of the unusual patterns in the detection curves generated here (Figure 4.5), for instance the smaller-than-expected declines in the High Monotone and Notch Monotone vocalization groups (*i.e.* Figure 4.5C and 4.5G). Future controlled experiments using observers of known hearing thresholds and ages, with exposure to a variety of bird vocalizations of known audiological characteristics, would help to elucidate the relative importance of hearing and non-hearing senescence effects, as well as the interaction between these processes and species-specific behaviours in the field.

To minimize the influence of hearing loss on the quality of bird surveys, Emlen and DeJong (1992) suggest that administrators test hearing abilities ahead of time and recommend the use of hearing aids where appropriate. Especially because hearing aids might not be practical or equivalent to normal hearing, we argue that administrators should also collect observer ages and information about hearing ability (once hearing aids are in place, if relevant) in order to make model-based corrections, for instance using the GAMM approach described here.

In both analyses, we controlled for many suspected observer-effects confounders by excluding data. For instance, we excluded observers aged 40–50 in the analysis of OBBA data to ensure reliable separation between 'young' and 'old' groups. We also required a minimum of 10 years of service on the BBS for an observer's data to be included, here to increase the likelihood that senescence effects occur for all observers, so that we might measure them. While these conservative approaches were appropriate for precisely determining the nature of observer senescence effects, they limit the quality of the real (simultaneous) population trends that can be inferred (Link and Sauer 1997*b*). Future analyses should also explore the sensitivity of the observer- and population-specific patterns we observed here to increasingly relaxed data subsetting rules. Similarly, having surveys collect observer age data in the future would obviate any future need to exclude cohorts of uncertain age, and allow models to make more-precise corrections than those we have built here.

GAM- and GAMM-based methods are relatively new to ecology (*e.g.* Fewster *et al.* 2000; Clarke *et al.* 2003; Flemming *et al.* 2010), but we have shown their usefulness for making corrections for continuous, nonlinear covariates (*i.e.* the changes in hearing ability with observer age). If these methods cannot be used, other, design-based remedies to the problem of observer senescence include simultaneous, independent sampling by multiple observers (Alldredge *et al.* 2006, but see Fitzpatrick *et al.* 2009), or the use of field recordings and more thorough and/or computer-aided *post hoc* interpretation (*e.g.* Campbell and Francis 2011). However any such protocol changes should aim to be consistent throughout the survey as a whole, they must be cost-effective, and ideally, they should not compromise the long-term integrity of the overall time series (*e.g.* Freeman *et al.* 2007).

At a minimum, asking older or noise-exposed observers who are at risk for detection errors to consider the possibility of any age-related impairments is an important step forward: as with any gradual physiological change, observers over age 50 may not recognize a growing, but significant personal impairment (A. G. Horn, pers. comm.), and awareness of this fact alone may lead to an increased degree of self-selection in terms of opting out of surveys. For instance, 75% of a sample of 253 Audubon Christmas Bird Count observers have indicated a desire to remove themselves from survey duties if such an impairment was recognized (Downes 2004). On the other hand, older birdwatchers are likely to be more experienced and consequently more adept at detecting a wide range of rare and common species. For effectively sampling entire species communities, this experience advantage may outweigh early deficiencies in the detection of certain species (Ramsey and Scott 1981), especially when there are multiple observers per sampling unit (but see Fitzpatrick *et al.* 2009).

In general, our study adds to the growing body of literature demonstrating systematic, long-term changes in BBS survey conditions (*e.g.* Betts *et al.* 2007; Griffith *et al.* 2010) that must be controlled for when estimating measures of population change. We have shown that observer age can be a significant handicap, and have illustrated some ways that survey designs and models might control for its effects. We hope that this research leads to improvements in long-term population trajectory inferences, without discouraging the invaluable contributions made by volunteers to worldwide ecological monitoring.

4.6 Acknowledgements

We thank T. Bishop from the Cornell Laboratory of Ornithology for assistance with the vocalization samples; W. Blanchard and I. Jonsen for statistical and computational advice; G. Breed and A. Horn for helpful discussions; and J. Bielawski, J. Mingrone and the Hutchings Lab for computing assistance. Thanks to the official sponsors of the Ontario Breeding Bird Atlas (Bird Studies Canada, Canadian Wildlife Service, Federation of Ontario Naturalists, Ontario Field Ornithologists, and Ontario Ministry of Natural Resources) for supplying Atlas data, and to the thousands of volunteer participants who gathered data for this project. We also thank the organizers, sponsors and volunteers of the North American Breeding Bird Survey for equally generous commitments of their time and skills.

4.7 Supplementary Material

4.7.1 Appendix 1: Vocalization Heterogeneity

By convention, sound intensities (power) are scored on the (logarithmic) decibel scale, which recognizes that human ears most readily distinguish changes in intensity along such an axis (Mayfield 1966). Converting a set of sound intensities to linear scales would tend to de-emphasize softer notes and highlight differences only among sounds of higher intensities. In our case, this approach was highly-appropriate for comparing vocalization variability in that it tended to downplay any background noises present on a given audio track and emphasize only the dominant singing and calling notes of a given species. Accordingly, to classify vocalizations into 'monotone' and 'heterogeneous' groups according to the variability of frequencies they contain, we first rescaled and linearized the log-scale decibel values within each power spectrum using the formula:

$$RelPower_i = 10^{(Power_i - Power_{max}) \cdot 0.1} \tag{4.1}$$

where $(Power_i - Power_{max})$ corresponds to the (negative) linear difference on the decibel scale between a given power value and the spectrum's maximum power value for $1, \ldots, i$ frequency bins. This function converts all decibel values to a scale from 0 to 1, where 1 equals the maximum power output, and it reflects linear-scale power differences (*i.e.* non-decibel values) between any given value and the maximum value. We then treated these power spectra as histograms and determined the standard deviation of these 'distributions' as a measure of their acoustic variability.

4.7.2 Appendix 2: Hierarchical Occupancy Model Structure

The occupancy component of the models for each species was specified as:

$$z_i \sim Bernoulli(\psi_i) \tag{4.2}$$

$$logit(\psi_i) = A_0 + A_1 \cdot \zeta_i \tag{4.3}$$

for $i = 1, \ldots, 1212$ atlas squares, and where z_i corresponds to the unobserved true occupancy state of a given (second-atlas) atlas square (*i.e.* 0 or 1), $P(z_i = 1) = \psi_i$

(the occupancy probability for atlas square i), and ζ_i is a dummy variable indicating detection/nondetection (*i.e.* 0, 1) of a species by any observer in square i in the first atlas (1981-1985). A_0 and A_1 are logit-scale intercept and first-year occupancy parameters. Data used to determine ζ_i were derived from a set of 1,325 total observers from the first atlas.

The detection component of the occupancy models for each species was specified as:

$$logit(p_{ij}) = \beta_1 \cdot \theta_{ij} + b_{obs_j} \tag{4.4}$$

$$b_{obs_j} = \beta_0 + \beta_2 \cdot Over50_j + \beta_3 \cdot Male_j + \epsilon_j \tag{4.5}$$

for $i = 1, \ldots, 1212$ atlas squares and $j = 1, \ldots, 296$ observers, and where p_{ij} is the detection probability at square i for observer j, θ_{ij} is the natural log of effort, in party-hours, at square i by observer j, β_1 is the effort effect, and b_{obs_j} describes the observer effects. Among these observer effects (equation 4.5), β_0 is an intercept term, β_2 is the age (over-50 *vs.* under-40) effect, β_3 is the effect of being male, and ϵ_j is mean-zero, normally-distributed error about the observer effect, with the uniformly-distributed variance of this error estimated from the data (see discussion of priors, below). $Over50_j$ and $Male_j$ are dummy variables (0 or 1) indicating whether an observer is over age 50 (*vs.* under age 40), and whether that observer is male (*vs.* female).

The occupancy and detection models are combined in the overall hierarchy, which incorporates observed detections Y_{ij}:

$$\mu_{ij} = z_i \cdot p_{ij} \tag{4.6}$$

$$Y_{ij} \sim Bin(N_{ij}, \mu_{ij}) \tag{4.7}$$

where Y_{ij}, the observed number of detections in square i for observer j is binomially distributed with probability of success μ_{ij} (the unconditional detection probability) for N_{ij} trials (*i.e.* the number of years during which an atlas square i was visited by observer j, which ranged from 2 to 5 detection-years).

All parameters in the hierarchical model (A_0, A_1, β_0, β_1, β_2, β_3) were assigned minimally-informative Bayesian priors suitable for logistic regression models, which in most cases need not estimate absolute values greater than 5 (Gelman *et al.* 2008). We specifically used normally-distributed priors of standard deviation 3.16 ($\sqrt{10}$), which, upon visual inspection of the density function, are distributed roughly similarly to the Cauchy prior of scale 2.5 recommended by Gelman *et al.* (2008, and see Figure 4.11) for this type of model – and which WinBUGS cannot directly simulate. We also considered other priors based upon the t-distribution that are also suggested by Gelman (2008); however, a sensitivity analysis showed that these priors (t_7 with scale parameters 2.5 and 10 for the predictors and intercept terms, respectively) had slower convergence rates, a narrower range of absolute parameter estimates, and lower effective sample sizes on average. This was a good indication that the normal priors were superior for our purposes.

The ϵ_j were assigned a normal, mean-zero prior with variance estimated from the data with a mean-zero, uniform prior (Gelman 2006) of standard deviation 10, which was again consistent with the range of parameter values expected in most logistic regressions (Gelman *et al.* 2008). We used enough iterations in WinBUGS to achieve convergence of 3 Markov chains (with a burn-in of one half of the total), requiring that Gelman-Rubin Rhat statistics for all parameters be less than or equal to 1.1 to infer convergence. We also verified the performance of this model structure using simulated datasets.

4.7.3 Appendix 3: Detailed Methods for Modeling Changes in BBS Count Data with Increasing Observer Age

To keep the more heavily-sampled species, observers or strata from having a disproportionate influence in our aggregated analysis, we modeled our BBS dataset over multiple stages using GAMMs. First, we modeled mean BBS counts for each species separately as overdispersed Poisson functions of both observer age and calendar year, correcting for differences among observers and survey routes as mean-zero, normally-distributed random intercepts. We used a cubic regression spline smooth term, chosen over thin-plate regression splines for computational efficiency reasons (Wood 2006), for each of the observer age and calendar year (*i.e.* population) effects, where the calendar year effects were smoothed separately for each stratum. The model structure for each species was as follows:

$$\log(y_{i(j)kl}) = f_1(\tau_{kl}) + f_2(l)_j + \theta_k + \lambda_{i(j)k} + \sigma_{i(j)kl} \tag{4.8}$$

for $i = 1, \ldots, I$ routes within stratum $j = 1, \ldots, J$, $k = 1, \ldots, K$ observers, and $l = 1, \ldots, L$ calendar years since 1969, and where $y_{i(j)kl}$ is the number of birds detected on a route i in stratum j by observer k during year l, $f_1()$ and $f_2()_j$ are cubic spline smooth functions estimating age effects across the whole survey and population-related effects for physiographic stratum j, respectively, τ_{kl} is the (minimum) age of observer k in year l, θ_k are mean-zero, normally-distributed random intercepts for each observer, $\lambda_{i(j)k}$ are mean-zero, normally-distributed random intercepts for each observer at a route-within-stratum, $\sigma_{i(j)kl}$ is mean-zero, normally-distributed overdispersion error, and where datapoints collected by a given observer were weighted according to the inverse of the number of routes conducted by that observer for the modeled species.

To properly recognize the changes in BBS counts predicted by the smooth function $f_1()$ in these models (Equation 4.8), we did not simply extract its values for the modeled range of observer ages, since this approach would ignore the uncertainty among the separate population-related smooth terms (estimated for each stratum; $f_2(l)_j$). Instead, working on the scale of the response variable, we defined species- and observer age-specific count predictions as the average of predictions for each relevant physiographic stratum. Calendar years were fixed at the midpoint of surveyed dates during predictions. We inferred the standard error about these averaged predictions, $\bar{\sigma}_{kl}$, as the square root of the mean of the variances of the initial predictions that were averaged.

We then built an ‘aggregating’ GAMM which generalized the predicted changes in BBS counts for each species with increasing observer age (*i.e.* the averages for each stratum and species produced in Equation 4.8) among each of eight vocalization frequency groups (*e.g.* ‘high monotone’, ‘notch heterogeneous’; discussed in Methods). In addition to generalizing the patterns of age-related count changes among species, this approach also ensured that each species contributed the same number of datapoints to the overall model. To convert the data to a common scale among all species,

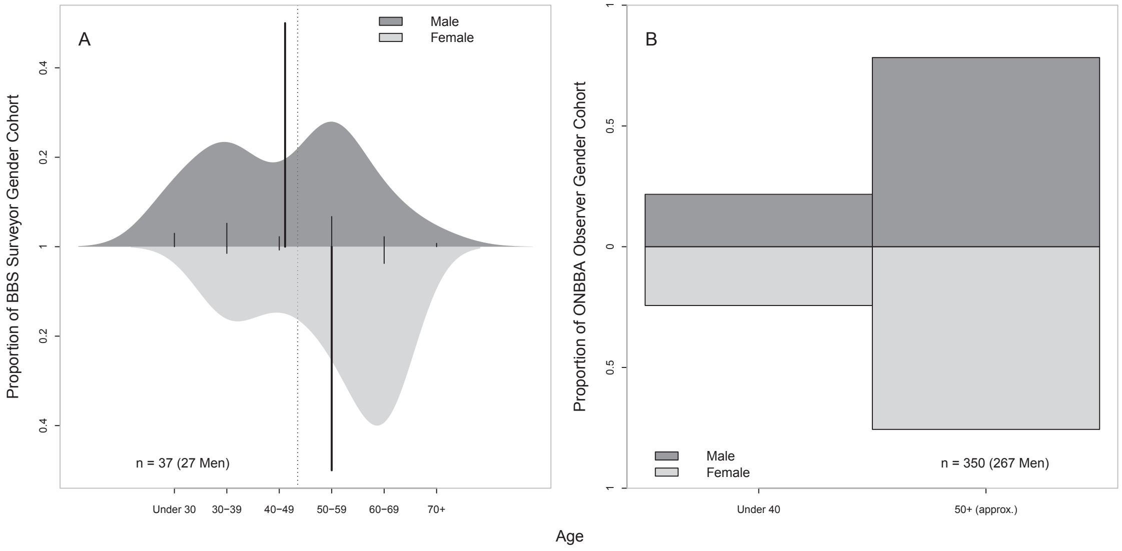

Figure 4.9. Age distributions among bird surveyor gender cohorts. Panel A shows a beanplot (Kampstra 2008) of the distribution of age ranges among a small sample of BBS observers, based upon demographic information collected by an unrelated internet-based survey of birdwatcher observer effects (Chapter 3). Tick mark lengths correspond to observer abundance at each age range; the dotted line is the overall mean, solid lines are group means. Panel B shows a barplot of the genders and estimated ages of those OBBA observers determined for the current study (not all observers participated over multiple years, and hence not all observers were modeled).

we used proportions of each species' maximum count as the (binomial) dependent variable in this model.

Similar to the single-species models, the aggregating GAMM used thin-plate regression spline smooth functions on observer age for each vocalization group, along with mean-zero, normally-distributed random intercepts for species. Each datapoint was weighted according to the inverse of its predicted coefficient of variation (*i.e.* $\frac{\hat{\mu}}{\hat{\sigma}}$). To provide a more useful interpretation of the species-independent changes in BBS counts with increasing observer age, final model predictions were then linearly rescaled relative to the values at observer-age 1 for each vocalization group. As in the detection probability analysis, we again validated the performance of our statistical approach by modeling simulated datasets of known population trajectories.

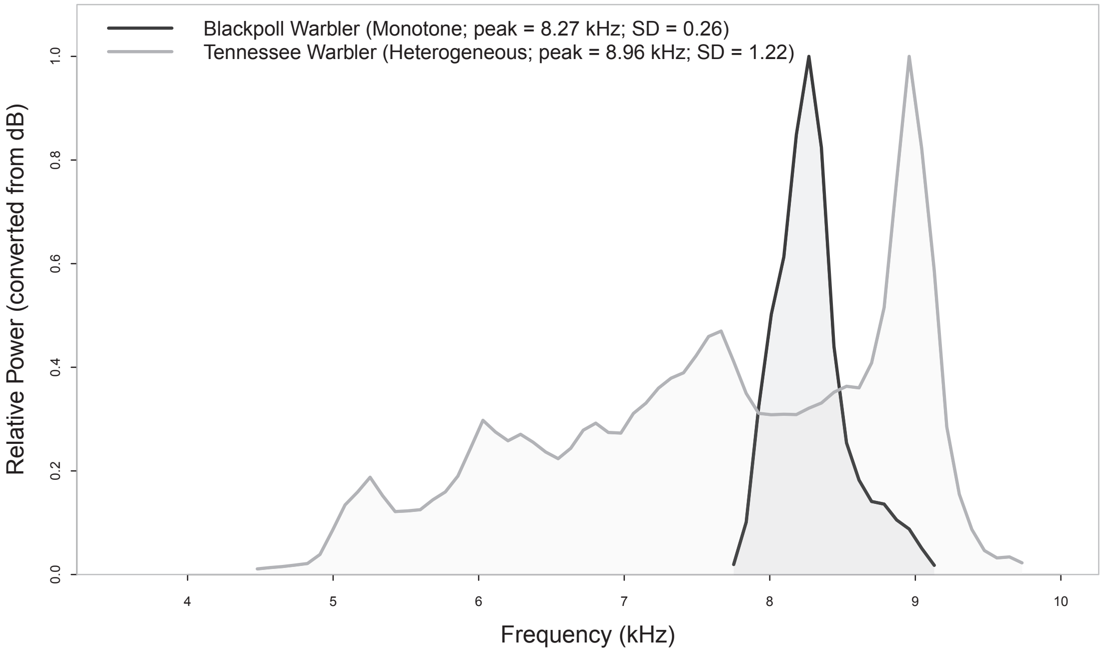

Figure 4.10. Examples of audiological power spectra corresponding to monotone (Blackpoll Warbler) and heterogeneous (Tennessee Warbler) vocalizations. The modified version displayed here presents the power as a linear-scale version of (normally log-scale) decibel values for each of a continuous range of frequency bins. Monotone vocalizations tend to feature a single or narrow range of frequencies, whereas heterogeneous vocalizations feature a wide range of sounds. Peak frequency values (kHz) and SD values (as a measure of heterogeneity) are listed for each species.

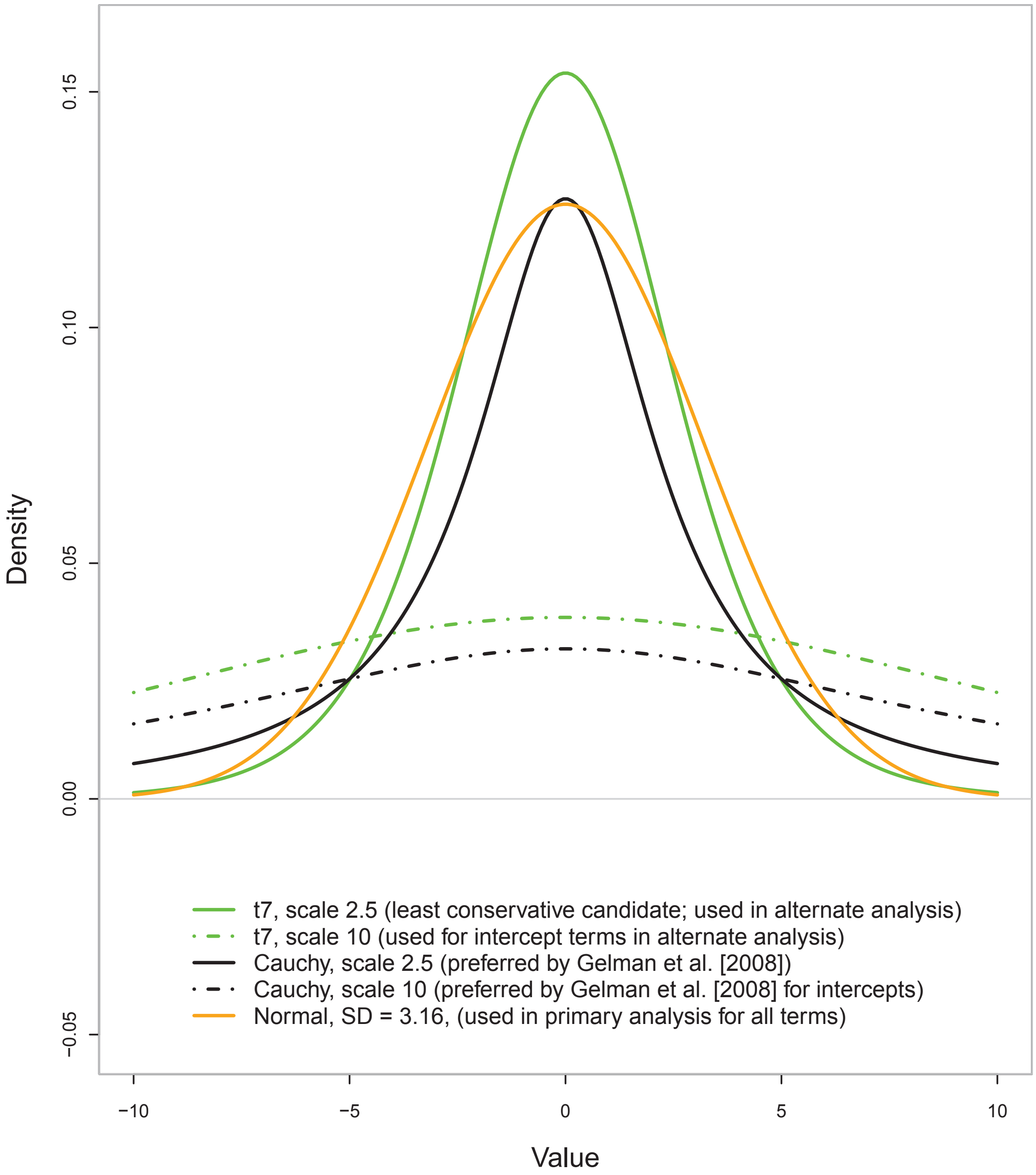

Figure 4.11. Probability density functions of selected prior distributions suggested or recommended by Gelman *et al.* (2008) for logistic Bayesian models, along with the N$(0,\sqrt{10})$ prior distribution used in the detectability change models here.

Chapter 5

Years-of-Service Effects in Long-Term Bird Survey Data

5.1 Abstract

Accurate population trend estimation from wildlife survey data must account for observer biases. Here, we built generalized additive mixed models to measure observer bias associated with long-term years of service in two volunteer bird surveys: the North American Breeding Bird Survey ('BBS') and the Audubon Christmas Bird Count ('CBC'). Using BBS data, we showed that as a single observer's years of service on a given survey route increase, all else being equal, expected bird counts increase by approximately $1.6 \pm 1.39\%$ over the initial 5 years, and then decline by $26.8 \pm 0.07\%$ by year 35 ($n = 33$ species). Among species not showing count increases during the initial 5 years ($n = 67$ species), the long-term effect is more pronounced, declining by $43.1 \pm 0.05\%$ by year 35. Using CBC data, we also show that, all else being equal, expected species richness on a given count circle increases by 14.4% (log-scale difference: 13.5 ± 3.2 units) as count circles are continuously surveyed by a party for 30 years. These patterns may reflect combinations of (i) growing initial familiarity with new survey sites; (ii) long-term sensory declines in individual BBS observers, but not among CBC survey groups; and (iii) a preference among CBC survey party members to meet or exceed a previous year's species richness counts. Simulated case studies highlight the value of accounting for these sources of error when estimating broad-scale ecological trends.

5.2 Introduction

Accurate population trend estimates are essential for wildlife population assessment and management. These estimates are typically derived for bird populations using field records from long-term count surveys such as the North American Breeding Bird Survey ('BBS'; Peterjohn 1994) and, less often, the Audubon Christmas

Bird Count ('CBC'; Dunn *et al.* 2005), both of which collect annual bird counts at predetermined sites. Part of the population trend estimation process involves accounting for confounding observer effects, namely differences in counts among and within unique observers. However, our knowledge of the processes driving these differences, and how to best control for their influences, is limited. In particular, there has been little focus on the influence of and correction for long-term changes within single observers over time ('years-of-service effects').

Analyses using BBS data account for differences in mean counts among unique observers by using random intercept terms (*i.e.* by assuming a normal distribution of observer abilities; Link and Sauer 1998, 2002). To partially account for years-of-service effects, these models also correct for lower-than-expected counts during an observer's first year of service on a given BBS route ("first-year effects"; Kendall *et al.* 1996). However, the models do not correct for additional within-observer changes occurring in later years such as a gradually-increasing familiarity with a survey site (*e.g.* Eglington *et al.* 2010), or declines in sensory ability due to age-related changes, including hearing loss (see Chapter 4).

Unlike analyses using BBS data, analyses using CBC data do not explicitly consider any years-of-service effects associated with individual observers, because counts in this survey consist of pooled observations made by groups of individuals. However, these survey parties have fairly consistent year-to-year membership (Butcher *et al.* 1990); hence, years-of-service effects may be present at the group level. For instance, party members might aim to match or exceed a previous year's recorded species richness (Bonta 2010, see also Jiguet 2009), and in doing so, spend more time at known 'hot-spots', or strategically distribute themselves according to past encounter histories. This could systematically inflate apparent species richness on the CBC over time, undermining our interpretation of the effects of broad-scale ecological processes (*e.g.* climate change; La Sorte *et al.* 2009).

Few studies have measured the magnitude of long-term years-of-service effects in surveys such as the BBS or CBC. Using BBS data, Link and Sauer (1998) modeled a 43% "diminuition of counts" among observers surveying Blue-gray Gnatcatchers (*Polioptila caerulea*) for more than 20 years, and in Chapter 4, we found that the observed counts of 60 species declined between 34% and 67% as minimum observer age (a variable closely related to years-of-service) increased over 39 years. To our knowledge, systematic changes in counts associated with years-of-service effects in the CBC have not been previously considered in published literature.

Our goal in this study was to test for long-term years-of-service effects in both the BBS and CBC. Compared to the previous research that used BBS data, we considered aggregated patterns from a much larger group of species (n = 100 [this study] *vs.* n = 65 [Chapter 4] and n = 1 [Link and Sauer 1998]). We also measured the influence of years-of-service on a given survey route, rather than minimum observer age (which, although closely-correlated, is not synonymous, because many observers survey multiple routes sequentially over their entire service history). This allowed us to better characterize the nature of the first-year effect (Kendall *et al.* 1996) in the context of later years-of-service effects such as aging-related sensory declines (see Chapter 4).

5.3 Methods

The relatively new technique of generalized additive mixed modeling (Wood 2006), available in package *mgcv* for the R statistical environment (R Development Core Team 2011) is well-suited for determining whether years-of-service effects might be important sources of error. Like locally-weighted scatterplot smoothing techniques ('loess'; James *et al.* 1996; Link and Sauer 1998), generalized additive models ('GAMs') and generalized additive mixed models ('GAMMs') fit continuous, smooth curves to predictor terms, displaying nonlinear patterns for any number of specified covariables. Using the *mgcv* package, the degree of smoothness of continuous GAMM functions is optimized and automatically selected as part of the fitting process (Flemming *et al.* 2010), which leads to increased objectivity and computational efficiency over loess fits, for which the degree of smoothness must be manually chosen. The 'significance' of a smooth function alone can only establish whether there is a nonzero pattern to the data – 'significant' GAMMs are not necessarily unidirectional (*i.e.* exclusively increasing or decreasing). Hence, GAM and GAMM smooths are visual instruments, and their shapes must always be examined in order to obtain a complete result.

GAMs and GAMMs also accommodate complex covariate structures because they can fit a simultaneous mixture of smooth curves alongside intercept terms and other parameters. This ‘semiparametric’ approach allows modelers to efficiently visualize the nonlinear influences of continuously-varying predictors such as an observer or survey party’s years of service, while also accounting for other fixed covariables such as survey location. Lastly, the random-effects components of GAMs and GAMMs consume fewer degrees of freedom when accounting for known grouping structures in the data than a fixed-effects strategy would, increasing the models’ predictive power.

We used GAMMs to test for changes in expected BBS counts with increasing years-of-service on a given survey route, and for changes in expected CBC species richness values over an increasing number of consecutive surveys, which we assumed reflected increasing party years-of-service. We focused on single-species trends with the BBS data, but on species richness trends for CBC data. Because of its dependence on species detection or nondetection, rather than on raw counts, species richness is a more appropriate dependent variable for the CBC’s less-controlled, variable-effort, areal-survey design (Dunn *et al.* 2005). Modeling species richness in this case also allowed us to measure the effect, if any, of the traditional goal for a CBC party to record a high number of species during a given survey (Preston 1958; Butcher *et al.* 1990; Bonta 2010). In each case (BBS and CBC), we then conducted a case study to illustrate the consequences of correcting for years of service when estimating long-term trends of species counts (BBS) or species richness (CBC), respectively.

5.3.1 Years-of-Service Effects (BBS)

The BBS is a North America-wide survey of avian abundance, conducted annually by skilled volunteers who make 3-minute point counts at each of 50 roadside locations along a predetermined, consistent survey route, recording all species seen and (or) heard at each location (Sauer and Link 2011). BBS data are available from 1966, however we used a subset of data from North American BBS routes surveyed between 1970 and 2007 in order to exclude known data quality problems with some early surveys of Canadian routes (http://ec.gc.ca/reom-mbs/default.asp?lang=En&n=E8974122-1, accessed on 25 March 2011). All data included unique observer identification codes. For each of 100 randomly-selected species present on the BBS (Appendix 1), we included single-species data sequences from single observers on routes featuring a minimum count of 5 birds per year, a minimum of 10 years of survey effort, and no gaps in survey coverage that were longer than 3 years. In our models, we corrected for the simultaneous effect of "real" population change by considering population trends at relevant Bystrak physiographic strata (Bystrak 1981). Each Bystrak stratum contains several survey routes, and so count trajectories shared among these routes (of the same stratum) indicate broader, regional patterns of population change independent of single-observer (route) effects. To ensure adequate replication of these regional population trajectories, we included only those strata with data from at least 3 survey routes.

We defined 'years of service' for each BBS data sequence for a given observer and survey route as the number of years since the first year of that sequence, accounting for the fact that some observers began participating in the BBS before 1970. To avoid confounding years-of-service effects with year-to-year population changes, we excluded data from Bystrak strata in which the pooled years-of-service among all participating observers had a Pearson correlation with calendar year that was greater than 0.7. An earlier study using simulated data showed that this approach adequately removed similar population effects (see Appendix 2 in Chapter 4).

Overall, these subsetting rules probably limited the realism of true population trajectories indicated by the analyses (sensu Link and Sauer 1997*a*). We nonetheless chose this approach in order to minimize errors in the data that could affect the shape of our covariate-of-interest (years of service).

Not all species are associated with significantly lower counts during the first year of an observer's service on the BBS (*e.g.* 35–48% of species; Kendall *et al.* 1996; Sauer and Link 2011). To account for these exceptions, we estimated the overall patterns of years-of-service effects separately among each of two species groups either showing or lacking first-year effects. To first identify these groups, we built overdispersed Poisson GAMMs similar to current modeling approaches (*e.g.* Link and Sauer 2002) predicting counts as a function of calendar year, with a covariable for the first-year effect, as well as random intercepts for observer and location (see Appendix 2 for details). Models testing for (linear, fixed-effects) first-year effects showed significantly ($p < 0.05$) lower counts during the first year for 32 of 100 species, and a significantly higher count for 1 species (Chihuahua Raven [*Corvus cryptoleucus*]; Appendix 1).

We considered these 33 species as the group to have shown significant first-year effects.

To next determine the continuous years-of-service effects for each first-year-effects group (showing or lacking first-year effects), we used GAMMs similar to the models of first-year effects (above), but here replaced the parameter corresponding to the first-year effect (‘η’, Appendix 2) with a cubic regression smooth function, $f_2(l - \tau_{i(j)k})$, where $\tau_{i(j)k}$ is the first year of service by observer k on route i within stratum j, and l is the survey year. This function showed how expected counts changed along a continuous range of years of observer service on a given survey route. Within each species-specific model, for each year of service, we predicted expected counts and errors (σ^2) for separate physiographic strata (all other variables being equal) and averaged these values among physiographic strata.

Using the species-specific (averaged) predicted counts, we then derived overdispersed binomial GAMM smooth functions describing the overall years-of-service effects separately for those species showing and lacking significant first-year-effects. In both cases, we re-scaled the single-species expected values to be relative to the value during the first year of service – thus re-expressing them as proportions – before building the model. Each of the two binomial GAMMs predicted expected proportional counts as a function of years of service, and included random intercepts for species. To account for different levels of uncertainty among each of the component species predictions used, these models also weighted individual datapoints according to the inverse of their coefficient of variation (*i.e.* $\frac{\mu}{\sigma}$).

We conducted a case study to illustrate the effect of correcting for long-term observer years of service in BBS population trajectory estimation, compared to older methods. We considered two older methods here: (1) ‘no correction’, where no years-of-service effects were modeled, and (2) ‘first-year correction’, where first-year effects, as described above, were modeled. Using data for the Vesper Sparrow (*Pooecetes gramineus*), a species with a large number of records and high mean counts, and which had an existing first-year-effect (Appendix 1), we compared estimated population trajectories (*i.e.* $f_1(l)_j$ in Equation 5.1, Appendix 2) under three correction scenarios (no correction, first-year correction, and “full” [continuous years-of-service] correction). For each of the three approaches, we used data from three selected geographic strata that showed uncorrected population trends that were apparently (i) increasing (Dissected Till Plains; Figure 5.3A), (ii) stable-to-decreasing (Till Plains; Figure 5.3B), or (iii) declining (Aspen Parklands; Figure 5.3C).

5.3.2 Years-of-Service Effects (CBC)

The CBC is an annual, area-based survey of winter birds conducted collectively by groups of volunteer observers of variable group size and ability who tally bird species abundances within a predetermined count circle 24.1 km in diameter (Francis *et al.* 2004; Dunn *et al.* 2005). We used a subset of data from all Canadian CBC count circles surveyed between 1961 and 2009 that were available in an electronic format.

Before calculating annual species richness values, we removed records not identified to species (*e.g.* 'Gull sp.') and all records of hybrid species. We also aggregated all subspecies, regional variants and 'morphs' or 'forms' (*e.g.* 'Dark-eyed [Oregon] Junco' *vs.* 'Dark-eyed Junco') into their parent taxa (sensu La Sorte and McKinney 2007). This approach reduced the chance that year-to-year inconsistencies in the data could artificially inflate or deflate species richness, for instance if one count circle participant identified multiple hybrids or morphs, while another participant in a previous year recorded only the parent category for members of the same resident population.

Butcher *et al.* (1990) describe the CBC as "often consist[ing] of an experienced core group of birders from year to year, with the same count compiler and the same party leaders." They add that "as a result, many CBC circles are covered in essentially the same way from year to year." We thus assumed that year-to-year continuity of surveys in the data implied that the same local coordinators and core participants were involved, and we only included count circles which were surveyed continuously for at least 10 years. We used the number of years over which a count circle had been continuously surveyed (its 'age') as a proxy for survey party years-of-service.

To account for the effects of real changes in species richness over time (*i.e.* with calendar year), we needed replicated time series for each survey location. To achieve this effect with the unreplicated CBC circles, we estimated species richness trajectories at a broader spatial scale, using count data from the individual count circles as replicates. We used the overlap between Bird Conservation Regions ('BCRs'; Sauer *et al.* 2003) and provincial boundaries as our broad-scale sampling units. As in the BBS analyses, to ensure adequate replication and separation of observer effects from real changes over time, we only considered BCR-provincial overlap regions containing at least 3 count circles and which had a Pearson correlation between pooled count circle 'ages' and calendar years that was less than 0.7 ($n = 6$).

Survey parties on the CBC are allowed to vary their survey effort from year to year, for instance by surveying for different lengths of time. Hence, effort is routinely accounted for as a group-level source of observer error (Link *et al.* 2006). Consistent with recent analyses of CBC data (*e.g.* Link and Sauer 1999; Link *et al.* 2006; Sauer *et al.* 2008), and for consistency among all years of data considered, we corrected for the effect of effort using overall party-hours as the indicator variable. We arbitrarily assumed a maximum realistic value for effort of 200 party-hours, which translates to an 11 h survey day with 18 people independently covering the circle – in our opinion, the upper limit of potential survey effort for most count circles. Consequently, we excluded data corresponding to effort values exceeding 200 h, which probably arise in most cases from incorrect recording or transcribing (sensu Peterson 1995). We also excluded data with missing or zero effort scores.

Using these data, we built a Poisson GAMM predicting annual species richness for a given count circle as a smooth function of years of service, correcting for the effects of calendar year and effort as additional smooth functions, and including a random intercept which controlled for differences in mean counts among each count circle. As a case study, we then graphically compared predicted species richness trajectories from this GAMM to a similar model built from the same data, but which excluded the correction for years of service (see Appendix 3 for details).

5.4 Results

5.4.1 Years-of-Service Effects (BBS)

Models testing for continuous years-of-service effects on the BBS showed significant ($p < 0.05$) nonlinear, long-term declines in expected counts for both those species showing significant first-year effects ($n = 33$), and those species showing no significant first-year effects ($n = 67$). Expected counts declined by more than 25% ($26.8 \pm 0.07\%$) in the group showing significant-first-year effects, and by more than 40% ($43.1 \pm 0.05\%$) in the group with no significant first-year-effects, respectively, as the observer reached 35 years of service (Figure 5.1). Among species showing first-year effects (Figure 5.1A), there was also a slight increase in counts in the early years of service (analogous to an 'extended' first-year effect) which peaked at $101.58 \pm 1.39\%$ of the original count after 5 years. Pointwise error estimates from the BBS models increased with years of service, reflecting the existence of increasingly fewer observers approaching 35 years of experience.

In the Vesper Sparrow case study, the first-year effect (in the first-year-effects model) corresponded to a 15.0% lower count compared to the average trajectory of subsequent years (log-scale difference: -0.162 ± 0.029; $p < 0.001$; Figure 5.2A). In the "full" model correcting for continuous years of service effects, estimated counts peaked at 5 years of service and were 4.6% higher than first-year counts (Figure 5.2B; log-scale difference 0.045 ± 0.097; SD obtained by simulation). Estimated counts steadily declined after that point, and at 35 years of service were 55.7% lower than the first year's value (Figure 5.2B; log-scale difference 0.783 ± 0.152 units; SD obtained by simulation).

Estimated count trajectories that corrected for first-year effects (Figures 5.3B, E, and H) were less-positive than uncorrected values (Figures 5.3A, D, and G). In contrast, count trajectories correcting for the full range of years of service were more positive than the uncorrected values, as well as than the values with first-year corrections (Figures 5.3C, F, and I). This shows the greater importance of the declining component of the years-of-service effect compared to the initial increasing component in the first 5 years (Figure 5.2B).

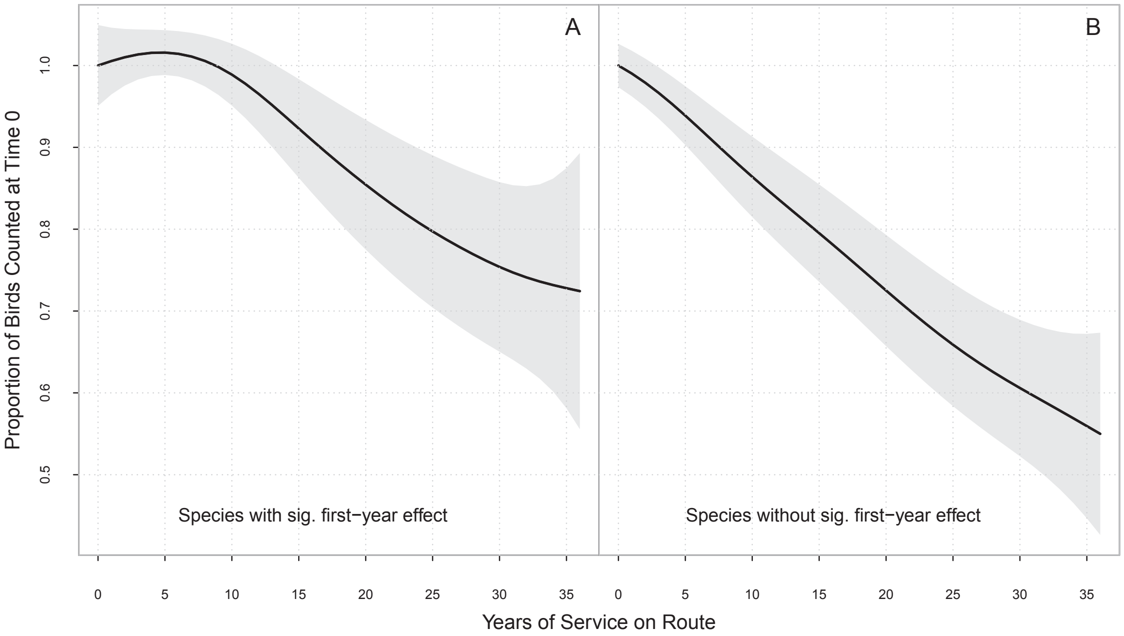

Figure 5.1. Overdispersed binomial GAMM smooth functions describing the proportion of expected species counts on BBS routes, relative to the counts made on the first year of service as a function of the number of years that an observer has surveyed the route. Panel A shows a geographic-stratum-averaged and uncertainty-weighted smooth function derived from count records of 33 species that showed significant first-year effects under a separate modeling approach. Panel B shows the smooth function of an identical model, but which was produced from count records of 67 species that did not show significant first-year effects. The peak in expected counts in panel A occurs at 5 years of service. Shaded areas are 95% pointwise confidence intervals.

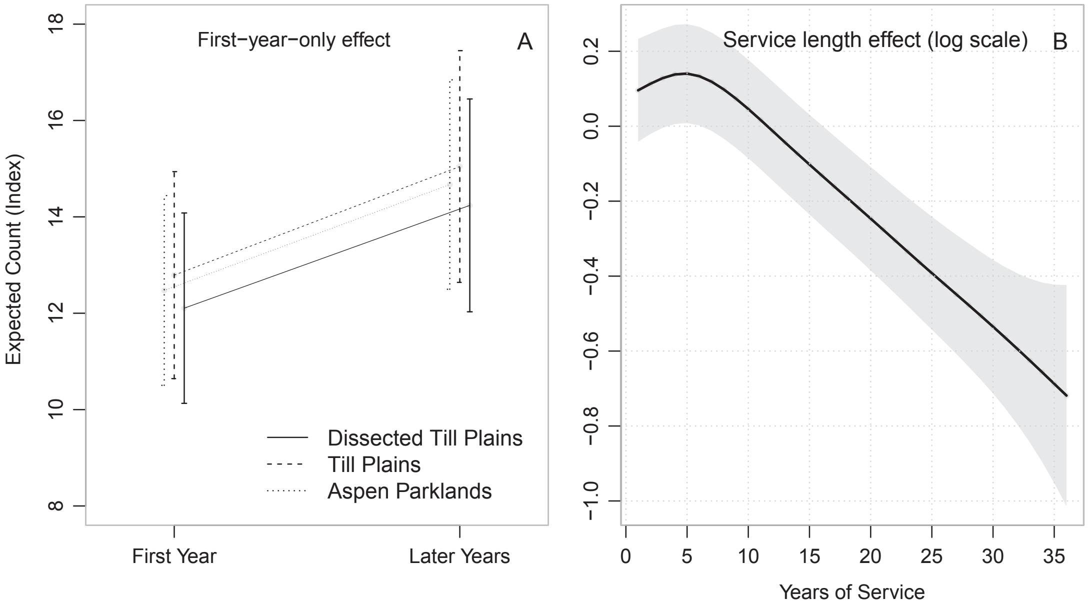

Figure 5.2. Years-of-service effects in the case study of Vesper Sparrow (*Pooecetes gramineus*) abundance estimates from BBS data. Panel A shows shows the first-year correction (log-scale) effect for each geographic stratum, modeled using a conventional technique (*i.e.* analogous to Link and Sauer 2002). Panel B shows the full service length correction for expected counts on the log scale, derived from the GAMM-based modeling approach presented here. All error bars and shaded areas are 95% pointwise confidence intervals. See Figure 5.3 for the effects of these corrections on the Vesper Sparrow abundance estimates.

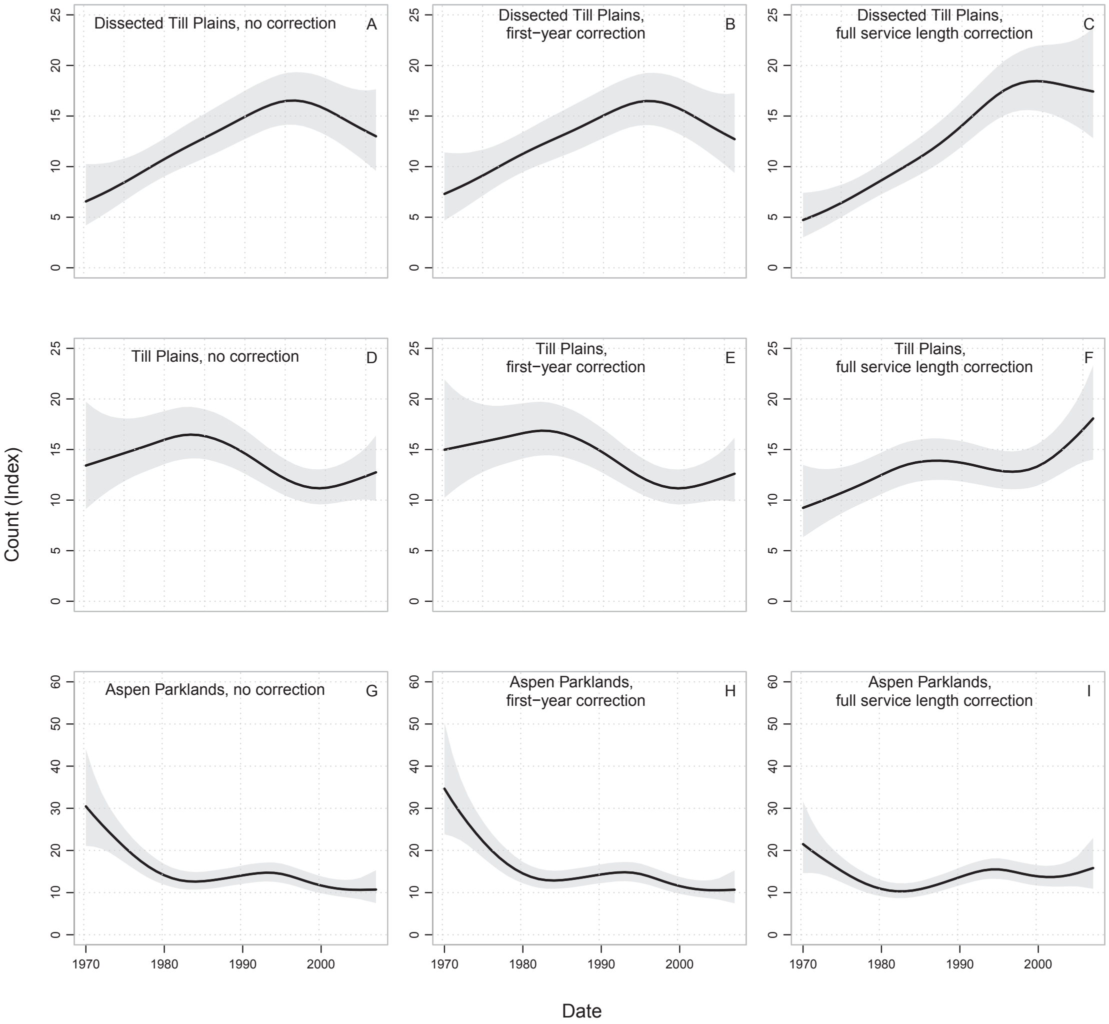

Figure 5.3. Case study of Vesper Sparrow (*Pooecetes gramineus*) years-of-service effects on abundance estimates from BBS data under three overdispersed Poisson GAMM modeling approaches that i) do not account for the effect of years of service on estimated population trajectories ("no correction"; panels A, D, G), ii) account for a years-of-service effect by considering first-year effects only ("first-year correction"; panels B, E, H), and iii) account for a years-of-service effect as a continuous smooth function ("full service-length correction"; panels C, F, I). Panels show the smoothed log-scale predictions of expected counts. All shaded areas are 95% pointwise confidence intervals. See Figure 5.2B to visualize the underlying years-of-service correction.

5.4.2 Years-of-Service Effects (CBC)

The model testing for continuous years-of-service effects on the CBC showed significantly nonzero, monotonic increases in observed species richness with increasing party years of service (count circle age; $p < 0001$; Figure 5.4G). All else being equal and averaged among BCR-provincial overlap regions, expected species richness increased by 14.4% (log-scale difference 13.5 ± 3.2 units; SD obtained by simulation) over 30 years of party attendance.

After correcting for years of service, four of the six geographic regions showed significantly nonzero, nonlinear patterns of change in species richness over time (Figures 5.4A through D: solid lines) and two showed no significant changes (Figures 5.4E and 5.4F: solid lines). In the (case-study) model that did not correct for years of service, all six geographic regions showed significantly nonzero, nonlinear species richness trajectories (Figures 5.4A through D: dashed lines).

Species richness trajectories tended to show shallower increases with calendar year when corrected for years-of-service effects (Figures 5.4A through F; solid lines), and richness values tended to be higher overall than uncorrected estimates (dashed lines).

5.5 Discussion

This research revealed a complex picture of observer years-of-service biases that could be affecting models of bird species richness and abundance. In the BBS models, among species showing first-year effects, we found a small increase in counts (approximately 101.58% of the first year's observations) developing between the first and fifth years of service on a route, and then a much larger subsequent decline – exceeding 25% – after 35 years (Figure 5.1A). This is consistent both with earlier research demonstrating a first-year-of-service effect (Kendall *et al.* 1996; Jiguet 2009), and with research showing observer senescence effects in both the BBS and an independent dataset (Link and Sauer 1998; and see Chapter 4). This pattern is also precisely what Bart *et al.* (2004*b*) speculated might tend to occur, arguing that growing familiarity with a given survey site probably leads to initial increases in expected counts, but that these gains are eventually superseded by normal, age-related losses

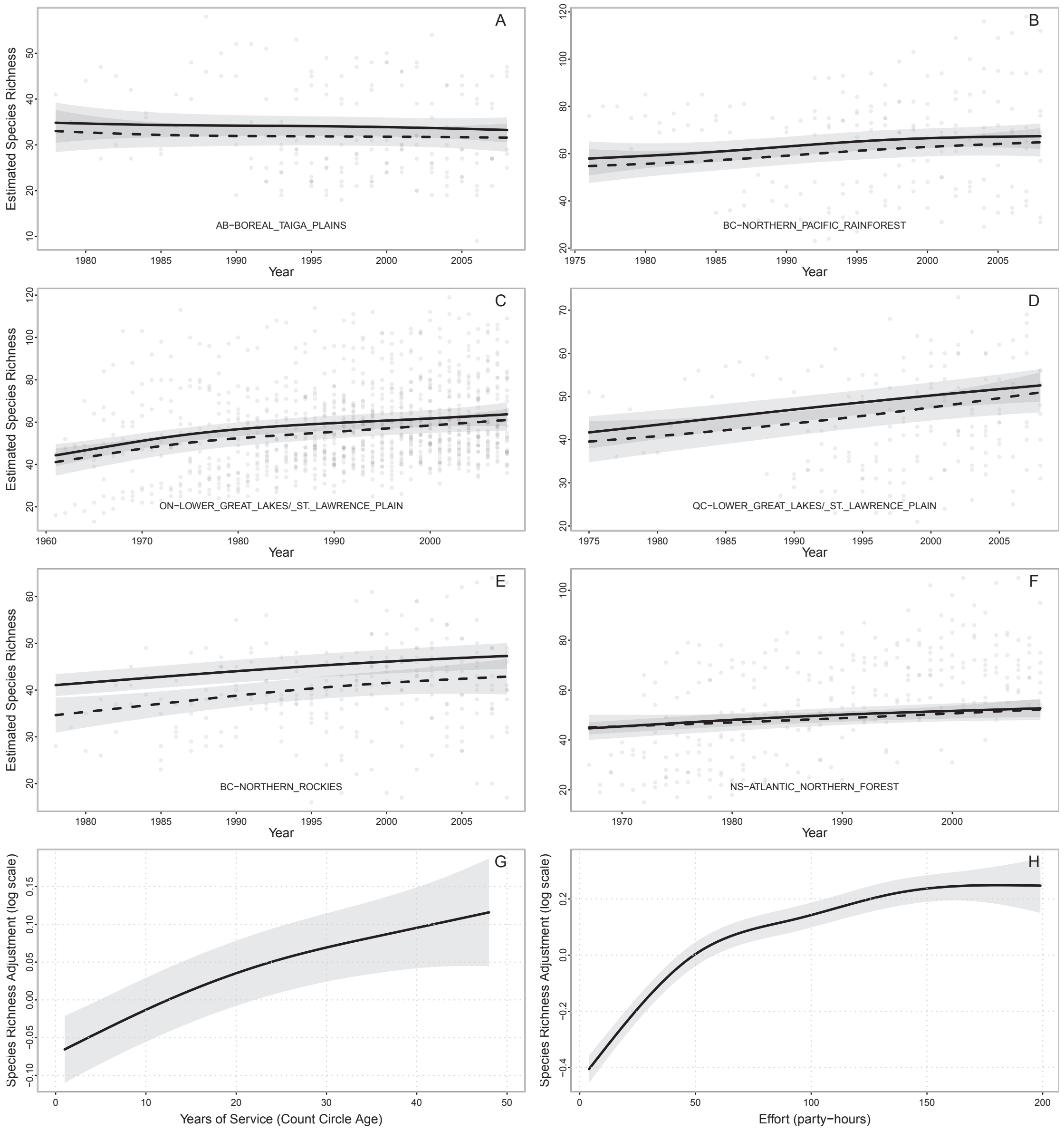

Figure 5.4. Trends in estimated Christmas Bird Count (CBC) species richness values for selected Bird Conservation Regions (Sauer *et al.* 2003) over time (panels A to F), plus the simultaneous, date-independent effects of years of service (count circle "age"; panel G) and survey effort (panel H). Points on panels A through F are raw richness values. Solid lines on these panels are GAMM estimates which correct for years of service; dashed lines are similar estimates which do not correct for years of service. Solid-line trends are significantly nonlinear (GAMM smooth term $p < 0.05$) in Panels A through D; all dashed-line trends are significantly nonlinear. Shaded areas are 95% pointwise confidence intervals.

of hearing and vision (*i.e.* sensory declines). Furthermore, the long-term decline in counts with observer service we observed here parallels results from models of British Breeding Bird Survey data (Eglington *et al.* 2010), in which the majority of models that accounted for a continuous linear effect of observer service predicted declining counts over time (*i.e.* a negative slope of the correction factor), which the authors – who did not consider the influence of sensory declines – found puzzling.

Our results also suggest that early increases in expected BBS counts seem to develop over periods longer than 1 year (*i.e.* 5 years), and that these increases are relatively small compared to the longer-term declines. Earlier research which recognized the existence of first-year-effects (Kendall *et al.* 1996) did not test for within-observer changes beyond year 1, thus, our continuous years-of-service approach also provides a more detailed picture of this process.

In the Vesper Sparrow case study (Figures 5.2 and 5.3), we showed how failing to correct for longer-term years of service effects, as is standard practise at present, led to more-negative BBS population trajectory estimates. This could lead to inappropriate management inferences and decisions (sensu Thomas and Martin 1996). Taken together, our data indicate a need to reconsider the years-of-service covariate structure of existing BBS population models, which we argue should correct for long-term years-of-service effects, rather than simply a first-year effect. In light of previous research showing the importance of more-direct measures of observer age (Chapter 4), future studies might also consider how to effectively control for the simultaneous influences of each of early and late years-of-service, and observer age. For instance, studies could evaluate whether modeling years-of-service could be adequate as a sole covariate, or whether a hybrid approach (*e.g.* by modeling early years-of-service effects and late observer-age effects as step functions) might be more appropriate.

We also identified significant years-of-service effects in CBC data. Here, we found a significant increase in species richness with increasing years of service (Figure 5.4E), independent of both calendar year (Figures 5.4A through F) and the effects of party effort (Figure 5.4G). The model predicted average species richness due to observer effects alone to increase 14.4% among count circles surveyed consistently for 30 years. This pattern may be a result of CBC participants attempting to maximize or outdo their previous years' species counts (Butcher *et al.* 1990), which is supported by observations that participants will often pre-survey count circles in advance of the formal CBC date (Preston 1958; Dunn *et al.* 2005). It is also qualitatively consistent with botanical studies showing species richness counts to be closely-related to surveyor experience and effort (Pautasso and McKinney 2007; Ahrends *et al.* 2011).

In the CBC case study, we demonstrated how, in practice, the trajectories of CBC species richness curves that correct for years-of-service effects (Figures 5.4A through F: solid lines) differ from those which do not (Figures 5.4A through F: dashed lines), in that correcting for years-of-service effects removed otherwise significantly nonlinear ($p < 0.05$) and positive trajectories of species richness in two of six geographic regions (Figures 5.4E and F). Similarly, the models illustrated how species richness values modeled with existing methods tend to be underestimates regardless of survey year, because most survey parties by definition do not have the highest amount of experience (Figures 5.4A through F; relative heights of solid *vs.* dashed lines). Collectively, this illustrates how failing to account for count circle 'age' can influence our determination of long-term patterns, in particular making long-term increases in species richness appear artificially steeper, and mean species richness estimates appear artificially lower.

The apparent increase of CBC species richness with increasing party years of service indicates a need for greater methodological and data controls, if CBC data collection is to serve scientific purposes beyond its primarily recreational intent. A recent, major review of the CBC (Francis *et al.* 2004) recommended establishing tightly-controlled sub-surveys within each count circle, with data recorded separately as a baseline to complement the larger, less-controlled count circle survey as a whole. We support this initiative; however, it has not yet been implemented. Similarly, this review recommended better-educating participants about the limitations of CBC data, especially when survey effort is not consistent. We also believe that administrators should more-strongly emphasize the value of adopting consistent search strategies from year to year, but we have not yet seen evidence that this is taking place. Lastly, we believe that in order to strike a balance between recreational freedom and methodological rigour on the CBC, model-based corrections – which do not necessarily require changes to survey protocols – are also useful. To facilitate making such model-based corrections, survey administrators should also begin collecting the identities of all count circle participants in order to establish accurate measures of a party's years of service.

Compared to decreasing long-term patterns on the BBS, the tendency for richness counts to increase with a CBC survey party's years of service, may reflect an 'advantage' of the group-based surveying approach over the solitary BBS surveying approach: in the CBC, older party members can work directly with younger survey participants, and this presumably compensates for any sensory changes that might bias the older observers' detection patterns over time (Stewart 1954; and see Chapter 4 and this paper).

Incidentally, our analysis of the CBC data indicated that Canadian winter species richness, corrected for the effects of years of service and survey effort, is nonetheless generally increasing over time in most ecoregions (Figures 5.4A through F). This echoes findings by La Sorte *et al.* (2009) and arguments by La Sorte and McKinney (2007) that Canadian circles, which tend to sit at the northern boundary of many species' ranges, might be particularly sensitive to the climatic and anthropogenic changes that can promote species colonization and extinction. Our research also suggests that correcting for years-of-service effects may lead to more precise estimates of this phenomenon.

One limitation of this study is its focus on elucidating the years-of-service effect outside the context of some other data quality issues affecting the BBS and CBC (*e.g.* missing data, short-term observers). For this reason, we believe that future methods research should consider the relative importance of these errors using larger, noisier datasets that the data subsets chosen here. Further studies into how observed species richness varies with observer experience in other competitive surveys such as eBird (Sullivan *et al.* 2009) might also be valuable. If years-of-service effects remain relatively important – which we expect to see, given the magnitude of the observed declines in this study (*i.e.* greater than 25–40% of initial counts on the BBS; Figure 5.1) – modelers should then determine how corrections for long-term years-of-service effects might be made both retrospectively to existing population trajectory estimates, and what modeling techniques might be most suitable for accounting for such processes in the future. Our work here, combined with previous studies of

GAMs (Link and Sauer 1997*a*; Fewster *et al.* 2000; Flemming *et al.* 2010) shows how additive models can be useful in such a role.

5.6 Acknowledgements

We thank the participants and coordinators of the BBS and CBC for their valuable, ongoing contributions to avian science and conservation. We also thank P. Blancher for helpful discussions and A. Horn and J. Flemming for providing comments on the manuscript.

5.7 Supplementary Material

5.7.1 Appendix 1: List of 100 Species Randomly Selected from BBS Routes From Which First-Year Effects Were Modeled as a Continuous (GAMM) Function. All significant first-year effects were negative (*i.e.* a lower-than-expected count during the first year of service) except for in the model of Chihuahua Raven (*Corvus cryptoleucus*) counts, where the first-year effect was positive.

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)	FY[a]
Acorn Woodpecker	*Melanerpes formicivorus*	503	3	32	30	25 ± 23	
Alder Flycatcher	*Empidonax alnorum*	2481	8	147	130	8.8 ± 9.1	*
American Crow	*Corvus brachyrhynchos*	16114	40	963	767	34.5 ± 25.6	*
American White Pelican	*Pelecanus erythrorhynchos*	64	1	4	4	23.6 ± 34.6	
Anna's Hummingbird	*Calypte anna*	204	1	12	13	5.7 ± 4.5	
Bachman's Sparrow	*Peucaea aestivalis*	245	2	16	13	8.3 ± 11.6	
Baird's Sparrow	*Ammodramus bairdii*	203	2	11	9	7.5 ± 8.8	
Band-tailed Pigeon	*Patagioenas fasciata*	571	4	31	33	6.9 ± 14.6	
Bay-breasted Warbler	*Dendroica castanea*	90	1	5	5	2.5 ± 2.4	
Black-billed Cuckoo	*Coccyzus erythropthalmus*	89	1	5	5	1.9 ± 2.2	

Continued on next page

[a] Significant first-year effect

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)	FY[a]
Black-crowned Night-Heron	*Nycticorax nycticorax*	78	1	4	3	1.9 ± 2.8	
Black-headed Grosbeak	*Pheucticus melanocephalus*	1534	9	91	76	14.4 ± 16.5	*
Black-throated Gray Warbler	*Dendroica nigrescens*	439	2	24	25	12.8 ± 17.1	
Blue-winged Warbler	*Vermivora cyanoptera*	384	2	22	17	4 ± 3.4	
Blue Jay	*Cyanocitta cristata*	13506	31	806	627	13.7 ± 11.5	*
Boat-tailed Grackle	*Quiscalus major*	269	2	15	11	28.1 ± 32	
Boreal Chickadee	*Poecile hudsonicus*	81	1	4	5	3.4 ± 2.7	
Brown-crested Flycatcher	*Myiarchus tyrannulus*	38	1	3	3	32.2 ± 18.9	
Brown Thrasher	*Toxostoma rufum*	7808	24	464	359	6.2 ± 5.3	*
Canada Goose	*Branta canadensis*	995	8	51	42	30.3 ± 75.7	
Canyon Towhee	*Melozone fusca*	190	3	12	10	9.4 ± 8.5	
Cassin's Kingbird	*Tyrannus vociferans*	187	2	13	11	17.3 ± 12.7	
Cerulean Warbler	*Dendroica cerule*	261	2	14	12	5.8 ± 5.9	
Chihuahuan Raven	*Corvus cryptoleucus*	108	1	7	4	15.2 ± 12.5	*
Cliff Swallow	*Petrochelidon pyrrhonota*	4176	23	251	222	47.3 ± 102.1	
Common Ground-Dove	*Columbina passerina*	629	4	37	25	7.7 ± 10.6	
Common Yellowthroat	*Geothlypis trichas*	13217	33	789	634	15.3 ± 12.6	*

Continued on next page

[a] Significant first-year effect

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)	FY[a]
Double-crested Cormorant	*Phalacrocorax auritus*	229	3	11	12	29.8 ± 63.2	
Eastern Kingbird	*Tyrannus tyrannus*	9081	26	532	415	7 ± 7.4	
Eastern Wood-Pewee	*Contopus virens*	7816	21	454	353	6.8 ± 5.2	*
Fish Crow	*Corvus ossifragus*	1514	4	92	66	9.7 ± 11	*
Forster's Tern	*Sterna forsteri*	59	1	4	4	5.2 ± 9.9	
Fox Sparrow	*Passerella iliaca*	320	4	19	18	17.3 ± 26.1	
Gilded Flicker	*Colaptes chrysoides*	60	1	4	3	8 ± 7.7	
Glossy Ibis	*Plegadis falcinellus*	40	1	3	3	52.3 ± 70.8	
Golden-crowned Sparrow	*Zonotrichia atricapilla*	52	1	4	3	40 ± 19	*
Golden-winged Warbler	*Vermivora chrysoptera*	60	1	3	3	3.3 ± 4.6	
Grace's Warbler	*Dendroica graciae*	86	1	6	5	9.7 ± 7.6	
Great Black-backed Gull	*Larus marinus*	165	1	8	9	8.8 ± 10.1	
Great Blue Heron	*Ardea herodias*	2248	12	131	108	3.9 ± 5.3	
Hairy Woodpecker	*Picoides villosus*	859	9	47	44	2.4 ± 2.1	
Hammond's Flycatcher	*Empidonax hammondii*	580	4	37	33	10.8 ± 10.1	*
Hermit Thrush	*Catharus guttatus*	3155	14	192	169	10 ± 11.7	
Hooded Oriole	*Icterus cucullatus*	46	1	3	3	5 ± 3.9	

Continued on next page

[a] Significant first-year effect

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)	FY[a]
Hooded Warbler	*Wilsonia citrina*	1757	8	104	73	5.2 ± 4.7	
Horned Grebe	*Podiceps grisegena*	116	1	5	3	2.4 ± 2.3	
House Finch	*Carpodacus mexicanus*	5385	26	310	246	15.7 ± 23.8	
House Wren	*Troglodytes aedon*	8482	27	494	416	12.3 ± 10.6	*
Hutton's Vireo	*Vireo huttoni*	340	2	19	17	3.8 ± 3.3	
Indigo Bunting	*Passerina cyanea*	9751	20	578	451	24.8 ± 19.6	*
Killdeer	*Charadrius vociferus*	12192	41	724	601	8.8 ± 8.7	
King Rail	*Rallus elegans*	41	1	3	3	5.4 ± 9.5	*
Ladder-backed Woodpecker	*Picoides scalaris*	195	3	14	14	5 ± 3.4	
Lark Bunting	*Calamospiza melanocorys*	641	3	41	35	127.2 ± 125.7	
Le Conte's Sparrow	*Ammodramus leconteii*	314	2	18	18	4.8 ± 5.3	*
Least Flycatcher	*Empidonax minimus*	3950	12	230	203	7.4 ± 6.3	*
Lesser Scaup	*Aythya affinis*	120	1	6	7	15.9 ± 33.4	*
Loggerhead Shrike	*Lanius ludovicianus*	1989	15	120	102	4.7 ± 5.2	
MacGillivray's Warbler	*Oporornis tolmiei*	880	5	53	47	7.1 ± 6.1	
Mallard	*Anas platyrhynchos*	4287	22	244	214	15 ± 34.4	
Marbled Godwit	*Limosa fedoa*	503	3	30	22	7.9 ± 11	

Continued on next page

[a] Significant first-year effect

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)	FY[a]
Mottled Duck	*Anas fulvigula*	102	1	7	7	12.2 ± 14.8	
Myrtle Warbler	*Dendroica coronata coronata*	1992	5	116	102	7.3 ± 7.3	*
Nashville Warbler	*Vermivora ruficapilla*	1936	6	113	101	12.1 ± 15.2	*
Northern Mockingbird	*Mimus polyglottos*	8604	31	527	415	25 ± 25.7	
Northern Pintail	*Anas acuta*	317	2	17	13	9.2 ± 15.1	
Northern Rough-winged Swallow	*Stelgidopteryx serripennis*	2549	15	140	113	4.7 ± 7.7	
Northern Waterthrush	*Parkesia noveboracensis*	662	3	40	38	4.7 ± 4.1	*
Nuttall's Woodpecker	*Picoides nuttallii*	267	1	17	18	6.8 ± 4.9	
Oregon Junco	*Junco hyemalis montanus*	1961	10	122	107	16.7 ± 14.8	
Palm Warbler	*Dendroica palmarum*	107	1	5	6	3.6 ± 3.6	
Phainopepla	*Phainopepla nitens*	140	2	9	7	10.2 ± 16.7	
Philadelphia Vireo	*Vireo philadelphicus*	205	2	13	11	3 ± 2.5	
Pileated Woodpecker	*Dryocopus pileatus*	2360	9	132	95	4 ± 4	
Pine Siskin	*Carduelis pinus*	1215	8	74	67	18.8 ± 24.2	
Pine Warbler	*Dendroica pinus*	2867	7	179	132	10.8 ± 11	*
Prothonotary Warbler	*Protonotaria citrea*	843	3	51	39	6.6 ± 10.1	*

Continued on next page

[a] Significant first-year effect

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)	FY[a]
Purple Finch	*Carpodacus purpureus*	2320	7	132	117	4.8 ± 5.2	
Purple Martin	*Progne subis*	4321	14	265	204	14.1 ± 20	
Red-breasted Sapsucker	*Sphyrapicus ruber*	64	1	4	3	8.7 ± 11.3	
Red-headed Woodpecker	*Melanerpes erythrocephalus*	2100	6	121	97	5.1 ± 5.6	
Red-shouldered Hawk	*Buteo lineatus*	420	2	24	15	3.3 ± 3.4	
Ruby-throated Hummingbird	*Archilochus colubris*	336	3	18	15	2.2 ± 2.3	
Savannah Sparrow	*Passerculus sandwichensis*	5926	23	355	304	18.5 ± 22.8	*
Say's Phoebe	*Sayornis saya*	477	5	31	25	4.8 ± 4.4	*
Scaled Quail	*Callipepla squamata*	254	2	16	9	12.9 ± 17.3	*
Seaside Sparrow	*Ammodramus maritimus*	81	1	5	4	10.9 ± 6.9	*
Slate-colored Junco	*Junco hyemalis hyemalis*	1494	6	88	78	7 ± 7.2	
Snowy Egret	*Egretta thula*	250	3	15	12	12.9 ± 27	
Song Sparrow	*Melospiza melodia*	11828	31	691	572	25.7 ± 21.2	
Sprague's Pipit	*Anthus spragueii*	168	2	8	9	6.6 ± 7.6	
Swainson's Thrush	*Catharus ustulatus*	2352	9	146	132	21.6 ± 23.8	*
Swamp Sparrow	*Melospiza georgiana*	1565	6	89	81	4.8 ± 4.5	*
Tree Swallow	*Tachycineta bicolor*	6344	22	370	314	10.6 ± 12.3	

Continued on next page

[a] Significant first-year effect

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)	FY[a]
Verdin	*Auriparus flaviceps*	232	3	15	12	15 ± 19.7	
Vesper Sparrow	*Pooecetes gramineus*	4147	17	240	202	21.5 ± 21.9	*
Western Meadowlark	*Sturnella neglecta*	4993	22	312	256	76.1 ± 91.6	*
White-crowned Sparrow	*Zonotrichia leucophrys*	397	2	25	24	15.9 ± 14.9	*
Yellow-bellied Sapsucker	*Sphyrapicus varius*	1439	4	80	71	5.9 ± 6.5	*
Yellow-breasted Chat	*Icteria virens*	4663	14	282	210	12.9 ± 11.4	*

[a] Significant first-year effect

5.7.2 Appendix 2: Details of the First Set of BBS Models (First-Year Effects Only)

GAMMs used to distinguish species with significant and non-significant first-year effects in BBS data followed the formulation:

$$\log(y_{i(j)kl}) = f_1(l)_j + \eta \cdot \mathbf{I}(i(j)k,l) + \theta_k + \lambda_{i(j)k} + \sigma_{i(j)kl} \tag{5.1}$$

for $i = 1,\ldots,I$ routes within stratum $j = 1,\ldots,J$, $k = 1,\ldots,K$ observers, and $l = 1,\ldots,L$ calendar years since 1969, and where $y_{i(j)kl}$ is the number of birds detected on a route i in stratum j by observer k during year l, plus a constant of 0.5 to better accommodate zero-counts (Sauer *et al.* 1996), $f_1(l)_j$ is a cubic spline smooth function estimating population-related effects (changes with calendar date) for physiographic stratum j, η is the first-year effect, $\mathbf{I}(i(j)k,l)$ is a dummy variable that equals 1 in an observer's first year of service on a route and 0 otherwise, θ_k are mean-zero, normally-distributed random intercepts for each observer, $\lambda_{i(j)k}$ are mean-zero, normally-distributed random intercepts for each observer at a route-within-stratum, $\sigma_{i(j)kl}$ is mean-zero, normally-distributed overdispersion error, and where datapoints collected by a given observer were weighted according to the inverse of the number of routes conducted by that observer for the modeled species within a given stratum.

In addition to using smooth functions in place of a large pool of random effects to account for nonlinear population trajectories, an important deviation from the model formula described in Link and Sauer (2002) is the inclusion of an observer effect (θ_k) alongside the nested observer-within-route effect ($\lambda_{i(j)k}$).

The case study of Vesper Sparrow (*Pooecetes gramineus*) population trends used the model in Equation 5.1 for the second scenario ('first-year correction'). For the first scenario ('no correction'), we modified this model by excluding the first-year effect variable (η). For the third scenario ('full correction') we modified this same model by replacing the first-year effect variable (η) with a cubic regression smooth function for years of service, $f_2(l - \tau_{i(j)k})$, where $\tau_{i(j)k}$ is the first year of service by observer k on route i within stratum j, and l is the survey year.

5.7.3 Appendix 3: Details of the CBC Data and Models

The principal Poisson GAMM for CBC data was specified as:

$$\log(y_{i(j)k}) = f_1(k - \tau_{i(j)}) + f_2(k)_j + f_3(\xi_{i(j)k}) + \theta_{i(j)} \tag{5.2}$$

for $i = 1, \ldots, I$ count circles within BCR-Province stratum $j = 1, \ldots, J$, and $k = 1, \ldots, K$ calendar years since 1969, and where $y_{i(j)k}$ is the number of birds detected on a route i in stratum j during year k, $f_1()$, $f_2()_j$ and $f_3()$ are tensor-plate, cubic spline smooth functions estimating (i) years-of-service effects, (ii) population-related effects for physiographic stratum j (here, with a ridge penalty allowing a zero-effect), and (iii) effort effects, respectively. $\tau_{i(j)}$ is the first year that a count circle i was surveyed, $\xi_{i(j)k}$ is effort in party-hours for each count circle i during year k, and $\theta_{i(j)}$ are mean-zero, normally-distributed random intercepts for each count circle.

We graphically compared species richness trajectory estimates predicted by this model, and by a second, nearly-identical alternative model built from the same data, but which did not include $f_1()$, the correction for years of service (Figures 5.4A through F).

Chapter 6

Re-evaluating the Interpretation of Apparent Longitudinal Changes in Observer Quality on the North American Breeding Bird Survey

6.1 Abstract

Past research measuring observer errors in the North American Breeding Bird Survey ('BBS') has shown that, averaged across their entire service histories, observers who began surveying a BBS route more recently have higher expected counts compared to observers whose 'start years' occurred earlier. Some studies have interpreted this result to mean that the overall 'quality' of new BBS participants must be systematically increasing over time. An alternative explanation for this pattern recognizes that observers with earlier start years tend to have been participating in the BBS for longer, and are thus, on average, older. Based upon known declines in observer detection ability with age, these older observers are therefore more likely to have lower average expected counts. Here, we show that observer start years and total years of service are negatively correlated. Accordingly, we also show that relative expected counts among observers increase with increasing start year and decrease with increasing years of service. Finally, we show that relative expected counts do not vary with start year in a group of observers with only short-term BBS service. Thus, while new BBS observers tend to count more birds than observers with less-recent start years, this pattern is not strong evidence that the year-to-year 'quality' of each successive cohort of observers has improved, but rather, it is more consistent with a decrease in the detection abilities of longer-serving observers within a given cohort over time.

6.2 Introduction

The current approach to modeling population trend data from the North American Breeding Bird Survey (“BBS”; Peterjohn 1994; Link and Sauer 2002) accounts for two types of observer effects. The first effect is a lower expected count during an observer’s first year of service on a survey route relative to subsequent years (‘η’; Kendall *et al.* 1996), while the second effect is a normally-distributed range of expected counts among unique observers visiting a particular survey route, relative to the mean for that route (‘Δb_j’ or ‘ω_j’ Sauer *et al.* 1994; Link and Sauer 2002).

Sauer *et al.* (1994) found that, for a given analysis year, observers have higher relative expected counts (*i.e.* higher Δb_j) if their first year of service on that route (‘start year’) is more recent than previous observers who they replaced. Subsequent research has revealed a similar trend with ω_j (Link *et al.* 2008). Authors have argued that this is evidence for a “trend over time of improving observer quality” (Sauer *et al.* 1994), an “increase in average birder skill over the past half-century” (Dunn *et al.* 2005), and a “change in the pool of observers” (Link *et al.* 2008). Sauer *et al.* (1994) attributed this pattern to a growing familiarity among volunteers with the BBS protocols and an increasingly-effective administrative system for disqualifying unskilled candidates.

We have previously shown that aging and long-serving BBS observers count fewer birds, presumably due to normal sensory declines that occur with age (Chapters 4 and 5). This has implications for the interpretation of the above patterns of Δb_j and ω_j. Here, we consider an alternative explanation for the increasing pattern of relative expected counts among observer cohorts with more-recent start years that takes these sensory declines into account. Because it recognizes that changes can occur within observers over the course of their service histories, this explanation does not imply that the ‘quality’ or ‘skill’ of new observers, defined in terms of expected counts, is improving from year to year. Instead, it only recognizes that newer (younger) observers tend to count more birds.

Under the proposed mechanism, we assume that the age at which any observer starts surveying the BBS tends to be similar throughout the BBS’s history. Given that observers who started surveying the BBS at earlier dates have longer service

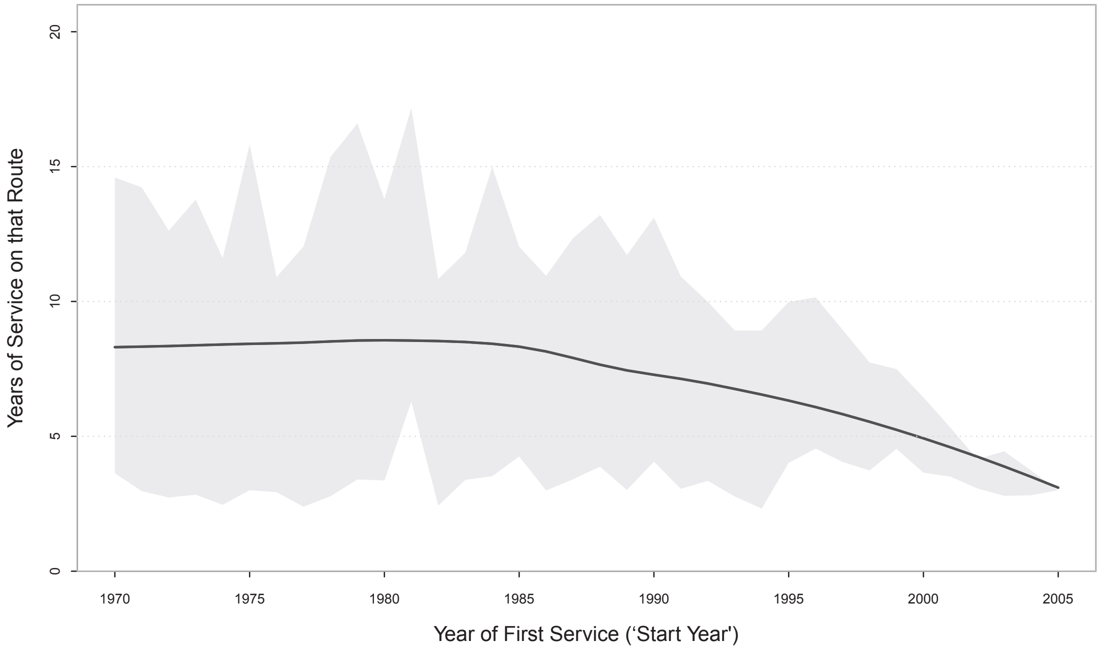

Figure 6.1. Relationship between years of service on a given BBS survey route (as of 2007) for a given observer and the date of that observer's first year of service on that route ('start year'). Years of service decline with more recent start years on any given survey route. Solid line is a loess smooth curve; shaded areas are one standard deviation of the mean. Data shown are identical to those used in later analyses.

histories on average ($r = -0.71$; Figure 6.1), we also assume that, on average, observers with earlier survey start years are older. Older (*i.e.* longer-serving) observers are more likely to have experienced age-related declines in their expected counts (see Chapters 4 and 5), and hence, their average expected counts should be lower. Taken together, this implies that the apparently increasing trend in average expected counts in more-recent start years occurs because, for a given point in time, observers with more-recent start years have had, by virtue of their being younger, less opportunity to experience age-related declines in their detection abilities.

We tested the above hypothesis by first calculating relative expected count scores, analogous to Δb_j, and measuring any covariation between these scores and an observer's years-of-service. We then used linear regression to show how the established, positive relationship between observers' relative expected count scores and their start years disappears in a dataset containing only observers with an identical, short service history, and hence, how this relationship depends upon there being older observers in the cohorts with earlier start years.

6.3 Methods

We derived relative expected count scores analogous to those used by Sauer *et al.* (1994) and Link *et al.* (2008) using overdispersed, Poisson generalized additive mixed models (‘GAMMs’; Wood 2006) of Canadian BBS count data (1970–2007) for 50 randomly-selected species (Appendix 1). These GAMMs were roughly equivalent to the existing standard for estimating BBS population trends from count data (*e.g.* Link and Sauer 2002), except they used continuous smooth functions in place of some highly-parameterized random effects in order to account for long-term population trends. Consistent with Link and Sauer (2002) but in contrast to Sauer *et al.* (1994), we did not calculate separate regressions for each species and survey route; instead, we used a single, hierarchical model for each species.

Each model incorporated normally-distributed random-effects intercepts corresponding to unique combinations of observer and survey route to account for among-observer variation in expected counts ($\lambda_{i(j)k}$; Appendix 2). To preserve the relative differences in these variables among observers with different service lengths, and to be consistent with models by Link and Sauer (2002), we did not account for the continuous effects of observer years of service as in Chapters 4 and 5, and instead only accounted for the known difference in counts between the first and second years of observer service (‘η’; Link and Sauer 2002). For computational efficiency, we also used BBS records collected exclusively by observers who surveyed Canadian routes for at least three years (range: 3–28 y). Additional modeling details can be found in Appendix 2.

As in the approach by Sauer *et al.* (1994), after deriving the models and their corresponding relative expected-count scores for each species (here, $\lambda_{i(j)k}$) we then calculated the deviation of each such term from the route-level means in each model. We refer to this new term, $\Delta\lambda_{i(j)k}$, henceforth as κ, and use it to describe relative differences in expected counts among observers for the rest of this study. To avoid issues of within-route autocorrelation, we then randomly selected one κ datapoint – representing a single observer’s relative expected count score – from each unique combination of species and route that was surveyed by at least two observers (sensu

Sauer *et al.* 1994).

We measured the correlation between an observer's length of service and his or her relative expected count score (κ) using simple linear regression. Based upon previous research (*e.g.* Sauer *et al.* 1994), and the negative correlation between start years and years-of-service (Figure 6.1), we expected that κ would decline with increasing years-of-service.

We then created a second linear regression to describe how the relative expected count scores (κ) varied with the observer's first year of service on that same route ('start year'), as in Sauer *et al.* (1994). We compared the slope of this regression to that of an identical linear regression built using only those observers showing five years of BBS service or less (regardless of their start year). We chose five years as a cut-off because previous research indicated that sensory declines might occur after this point among BBS observers, assuming a consistent age of first service (see Chapter 5).

If systematic differences in relative expected counts among observers with different start years are largely the result of years-of-service effects (*i.e.* aging), we expected that whereas the κ scores would increase as start years became more recent among observers with variable service history lengths (as in Sauer *et al.* [1994] and Link and Sauer [2008]), in the regression which used exclusively short-term observers, this trend would not differ significantly from zero. This is because no systematic, age-related differences in detection ability should be present among the different start-year cohorts if the corresponding observers' service histories (and hence, ages) are roughly equivalent.

6.4 Results

The linear regression predicting relative expected count scores (κ) as a function of observer years of service was significantly negative ($\beta_{Service} = -0.0033$, $p = 0.010$, $df = 2629$; Figure 6.2). This regression was not sensitive to the influence of outliers; removing those individuals with more than 20 years of service post hoc ($n = 58$ of 2631 records) did not affect its directionality or significance ($\beta_{Service} = -0.0038$, $p = 0.010$, $df = 2571$).

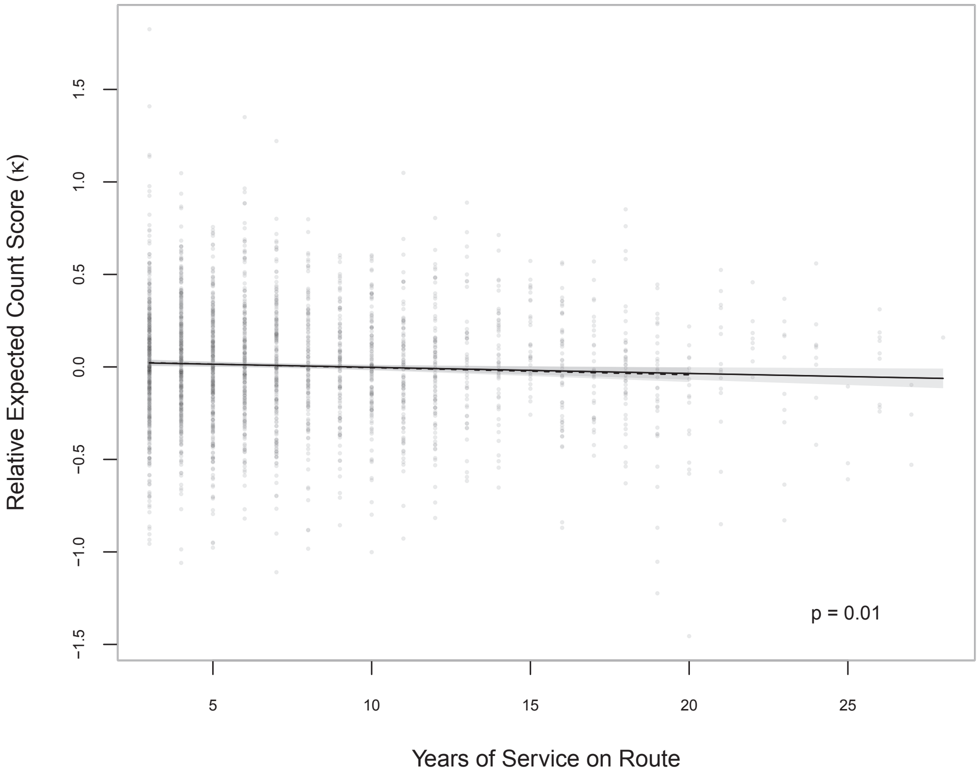

Figure 6.2. Relative expected count scores among BBS observers (κ), derived using an approach similar to that in Sauer *et al.* (1994) using BBS data from 50 randomly-selected species, randomly sampling one such score per species-route combination (to avoid autocorrelation within routes), and plotting these values as a function of the observer's number of years of service on a BBS route. Both lines are significantly negative linear regressions: the solid line ($p = 0.010$) is a regression on all datapoints; the dashed line ($p = 0.010$) illustrates that the significance is not dependent on effects from outlying observers (*i.e.* with more than 20 years of service; $n = 58$ of 2631 records). Shaded areas are 95% confidence intervals.

The linear regression predicting relative expected count scores (κ) as a function of observer start year indicated a small, but significant increase in expected counts with increasing (*i.e.* more-recent) start year ($\beta_{startYear} = 0.0030$, $p < 0.001$, $df = 2629$, Figure 6.3A). Restricting the analysis to those observers with 5 years of service or less, an identical regression was no longer significantly different from zero ($\beta_{startYear} = 0.0008$, $p = 0.407$, $df = 873$; Figure 6.3B).

To test whether the slopes of the latter two regressions of κ on observer start year were significantly different from each other, we next built an interaction model, specified as

$$\kappa_{ij} = (\beta_{startYear} + \beta_{int} \cdot \pi(i)) \cdot t + \beta_{short} \cdot \pi(i) + \theta + \epsilon_{ij} \tag{6.1}$$

where κ_{ij} is the relative expected count score described above for observer $i = 1, \ldots, I$ on BBS route $j = 1, \ldots, J$; $\pi(i)$ is a dummy variable which equals 1 if observer i served a total of 5 years or less during his or her entire service history on the BBS (any routes), and 0 otherwise; t is the first year of service; θ is an intercept term; and ϵ_{ij} is residual error. In this model, consistent with earlier results, β_{int} was significantly negative ($p = 0.033$), indicating that the slopes were indeed significantly different among observers with short (≤ 5 y) and long (> 5 y) BBS service histories.

6.5 Discussion

Sauer *et al.* (1994) described an apparent improvement in mean observer 'quality' for observers beginning service on a BBS route more recently. They attribute this pattern of increasing average expected counts among newer observers to their being more-familiar with the BBS survey protocol than earlier BBS participants, and to BBS administrators using increasingly effective techniques for selecting observers. Our results point to a different explanation. Specifically, the apparent increase in relative expected counts among observers newer to a BBS route may result from their having had fewer years of service, and so less opportunity to age and to experience sensory declines (*i.e.* Chapter 4; Figures 6.2 and 6.3). This implicitly recognizes that the expected counts of a given observer cohort are not constant over time, and if recruitment demographics remain consistent, that newer cohorts will always tend to

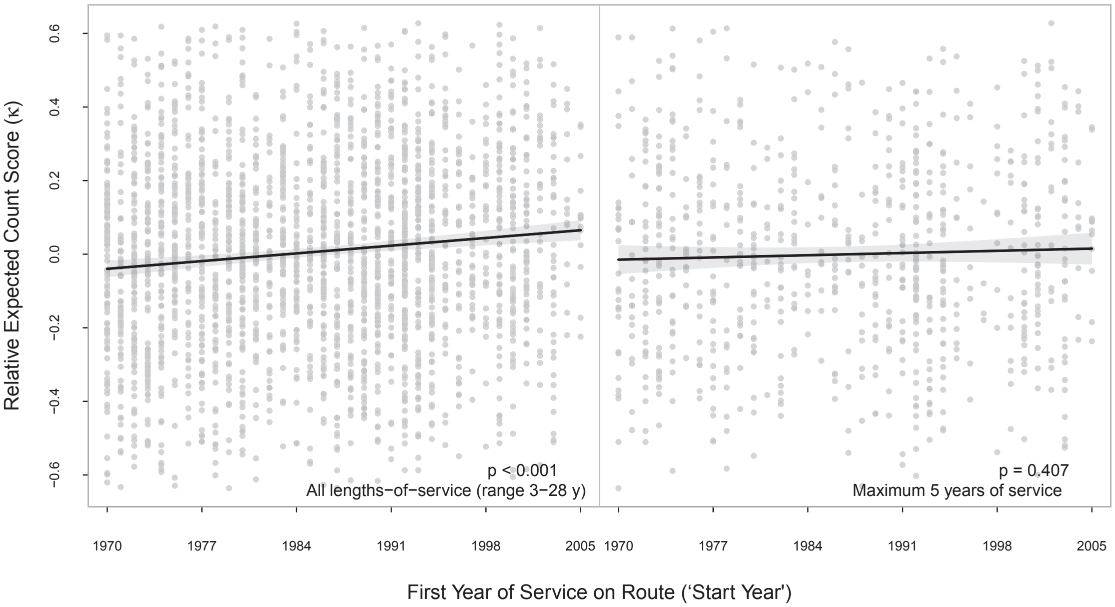

Figure 6.3. Relative expected count scores among BBS observers (κ), derived using an approach similar to that in Sauer *et al.* (1994) using BBS data from 50 randomly-selected species, randomly sampling one such score per species-route combination (to avoid autocorrelation within routes), and plotting these values as a function of the observer's first year of service on the BBS. Panel A includes all potential start years, whereas Panel B is restricted to observers having a maximum 5 years of service. Raw data (points) otherwise included in the analysis are clipped at the 2.5% and 97.5% quantiles in this graphic to avoid having overly-broad y-axis limits, and to better illustrate differences in the slopes of the linear regressions (solid lines). P-values for the significantly positive (Panel A) and non-significant (Panel B) regressions are listed. Shaded areas are 95% confidence intervals.

count more birds because of their youth. Their expected counts will in turn decline relative to even-newer observers as time passes. Hence, the apparent "trend over time of improving observer quality" (Sauer *et al.* 1994) is more likely a "tendency for newer (younger) arrivals to the BBS on any given year to count more birds than longer-serving (older) participants."

Despite its non-significance, the regression of relative expected count scores with increasing start year among observers with five or fewer years of service still has a positive slope estimate (Figure 6.3B). While the sample of data used for this regression ($n = 875$) is approximately one third of the data used in the full analysis ($n = 2361$; Figure 6.3A), it is still substantial, suggesting that the pattern we observed is robust. This is further supported by the interaction model which showed a significant difference between the non-significant relationship in this dataset, and the significantly increasing pattern found when observers of all service lengths are modeled (the 'full analysis'). Although we cannot say that processes such as better selection techniques by BBS administrators (sensu Sauer *et al.* 1994) are not also influencing long-term patterns of relative expected counts among observers, using these data, we can confidently say that any such processes are much less-important compared to years-of-service and observer age effects.

We assumed that the average age at which a BBS observer tends to start surveying a route has been consistent through time. If this assumption is not correct, some cohorts of observers with 5 years of experience or less might have already experienced age-related declines. If this pattern differs systematically among cohorts (*e.g.* with some cohorts being much older or much younger than adjacent ones), any long-term trends in relative expected counts among observers – or the absence of such trends – would not be consistent with our proposed mechanism. Unfortunately, we do not have true ages of our sample of BBS observers, and instead can only consider indirect information on demographics of birdwatchers in general. For instance, Wiedner and Kerlinger (1990) found that the average age of birdwatchers participating in the Audubon Christmas Bird Count in the United States in 1990 was 47, whereas La Rouche (2001) reported that the average age of American "birders" (people who had travelled more than a mile from home to see birds, or who had tried to identify birds around their homes) in 2001 was 49, and Carver (2009) reported an average age of 50 in 2006. If these data are indirect indicators of demographic trends among new BBS observers in particular, then they do not support our assumption of consistent ages. However, our analysis is concerned with the ages of observers during their start years, and not the average ages of birdwatchers of all levels of experience, as these data show. Furthermore, BBS observers tend to be highly-experienced, and often are professional ornithologists (Ziolkowski Jr. and Pardieck 2006), whereas these surveys poll the birdwatching public in general. Hence, these results are not necessarily relevant to our analysis.

Our new interpretation of systematic differences in relative expected counts among observer cohorts of different start years is not immediately relevant to the success of associated population modeling, so long as inter-observer variations are taken into account. In other words, the explanation for differences in among-observer expected count coefficients over time is less important than the fact that such coefficients be used in predictive models. However, this research is a reminder that significant within-observer changes can take place over time, and that these changes can affect our interpretations of statistical patterns. The results also make an important philosophical contribution which might be appreciated by more-senior BBS participants. Here, we found no evidence supporting the notion that observers beginning their service on the BBS during the 1970s and 1980s were any less-skilled at the time than are observers beginning their service today.

6.6 Acknowledgements

We thank the participants and coordinators of the BBS for their valuable, ongoing contributions to avian science and conservation. We also thank J. Flemming and A. Horn for providing helpful comments on the manuscript.

6.7 Supplementary Material

6.7.1 Appendix 1: List of 50 Species Randomly Selected from BBS Routes From Which Trends in Relative Expected Count Scores Among Observers (sensu Sauer *et al.* 1994) Were Calculated.

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)
Acadian Flycatcher	*Empidonax virescens*	2664	12	200	178	8.5 ± 6.8
American Bittern	*Botaurus lentiginosus*	631	7	55	55	5.3 ± 6.9
Ash-throated Flycatcher	*Myiarchus cinerascens*	2773	17	251	252	16.3 ± 15.2
Bachman's Sparrow	*Peucaea aestivalis*	342	3	31	31	9.4 ± 11.2
Boat-tailed Grackle	*Quiscalus major*	795	5	68	69	37.1 ± 72.2
Broad-tailed Hummingbird	*Selasphorus platycercus*	612	2	67	62	12.8 ± 11.5
Bronzed Cowbird	*Molothrus aeneus*	198	3	22	28	15.4 ± 29.2
Brown-crested Flycatcher	*Myiarchus tyrannulus*	225	3	23	24	15.9 ± 17.2
California Quail	*Callipepla californica*	2176	12	183	219	14.4 ± 15.6
Chestnut-sided Warbler	*Dendroica pensylvanica*	4932	11	384	414	12.5 ± 10.3
Common Moorhen	*Gallinula chloropus*	208	2	20	19	12.7 ± 19.4
Common Tern	*Sterna hirundo*	166	5	15	23	7.9 ± 10.9
Double-crested Cormorant	*Phalacrocorax auritus*	1237	17	101	106	14.3 ± 38.3

Continued on next page

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)
Evening Grosbeak	*Coccothraustes vespertinus*	2746	13	221	222	10.8 ± 35.8
Gilded Flicker	*Colaptes chrysoides*	118	2	13	15	9.8 ± 8.4
Golden-crowned Sparrow	*Zonotrichia atricapilla*	140	3	15	14	37.2 ± 22.4
Golden-winged Warbler	*Vermivora chrysoptera*	121	2	10	9	6.4 ± 4
Gray-headed Junco	*Junco hyemalis caniceps*	520	2	57	55	14.2 ± 14.8
Great Crested Flycatcher	*Myiarchus crinitus*	10722	32	808	770	8.5 ± 7.5
Greater Roadrunner	*Geococcyx californianus*	221	4	22	18	5.6 ± 5.3
Great-tailed Grackle	*Quiscalus mexicanus*	1698	13	145	151	39 ± 162.3
Hermit Thrush	*Catharus guttatus*	5804	24	547	535	12.6 ± 13.7
Least Tern	*Sternula antillarum*	290	6	24	26	22.4 ± 43.6
MacGillivray's Warbler	*Oporornis tolmiei*	1856	11	171	177	9.9 ± 8.1
McCown's Longspur	*Rhynchophanes mccownii*	327	5	37	37	17.7 ± 48.1
Mexican Jay	*Aphelocoma wollweberi*	79	1	7	8	30.9 ± 27.3
Northern Pintail	*Anas acuta*	1493	11	128	126	8.8 ± 24.3
Northwestern Crow	*Corvus caurinus*	383	3	32	35	37.2 ± 38.1
Orchard Oriole	*Icterus spurius*	5766	23	454	376	8.3 ± 7.2
Oregon Junco	*Junco hyemalis montanus*	3658	13	307	330	18.1 ± 16
Ovenbird	*Seiurus aurocapillus*	9161	22	722	746	18.4 ± 14.2

Continued on next page

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)
Painted Bunting	*Passerina ciris*	1870	12	163	180	15.3 ± 14.3
Pine Warbler	*Dendroica pinus*	4795	14	392	375	12.3 ± 11.6
Red-eyed Vireo	*Vireo olivaceus*	15935	31	1244	1250	23 ± 21.4
Ruby-crowned Kinglet	*Regulus calendula*	3262	13	335	322	14 ± 13.5
Ruddy Duck	*Oxyura jamaicensis*	778	8	65	70	7 ± 13
Rufous Hummingbird	*Selasphorus rufus*	552	3	49	53	6.4 ± 5.2
Savannah Sparrow	*Passerculus sandwichensis*	10284	34	874	932	21.5 ± 25.7
Say's Phoebe	*Sayornis saya*	756	9	86	77	7.1 ± 6.7
Sedge Wren	*Cistothorus platensis*	1140	9	92	89	8.1 ± 8.3
Spotted Sandpiper	*Actitis macularia*	392	5	34	40	7.6 ± 5.6
Tricolored Blackbird	*Agelaius tricolor*	365	4	32	33	122.4 ± 375.9
Varied Thrush	*Ixoreus naevius*	1306	9	136	137	19.9 ± 20.1
Virginia's Warbler	*Oreothlypis virginiae*	321	3	33	28	11.6 ± 17
Western Bluebird	*Sialia mexicana*	1020	7	88	102	9.6 ± 9.2
Western Gull	*Larus occidentalis*	164	2	13	15	29 ± 40.2
White-crowned Sparrow	*Zonotrichia leucophrys*	2003	17	202	219	18.8 ± 21.3
White-faced Ibis	*Plegadis chihi*	379	5	28	35	63.5 ± 141.3
White-throated Swift	*Aeronautes saxatalis*	465	9	49	50	10.8 ± 22.7

Continued on next page

Appendix 1, continued

Species	Scientific Name	Records	Strata	Routes	Observers	Mean Count (± SD)
Wilson's Warbler	*Wilsonia pusilla*	1928	14	200	198	13.8 ± 16

6.7.2 Appendix 2: BBS models

GAMMs used to determine relative among-observer expected count scores for each species followed the formulation:

$$\log(y_{i(j)kl}) = f_1(l)_j + \eta \cdot \mathbf{I}(i(j)k, l) + \lambda_{i(j)k} + \sigma_{i(j)kl} \tag{6.1}$$

for $i = 1, \ldots, I$ routes within Bystrak (geographic) stratum $j = 1, \ldots, J$, $k = 1, \ldots, K$ observers, and $l = 1, \ldots, L$ calendar years since 1969, and where $y_{i(j)kl}$ is the number of birds detected on a route i in stratum j by observer k during year l, plus a constant of 0.5 to better accommodate zero-counts (Sauer *et al.* 1996), $f_1(l)_j$ is a cubic spline smooth function estimating population-related effects (changes with calendar date) for physiographic stratum j, η is the first-year effect, $\mathbf{I}(i(j)k, l)$ is a dummy variable that equals 1 in an observer's first year of service on a route and 0 otherwise, $\lambda_{i(j)k}$ are mean-zero, normally-distributed random intercepts for each observer at a route-within-stratum, $\sigma_{i(j)kl}$ is mean-zero, normally-distributed overdispersion error, and where datapoints collected by a given observer were weighted according to the inverse of the number of routes conducted by that observer for the modeled species within a given stratum.

Chapter 7

Discussion

Compared to conventional data collection strategies that use students, scientists and paid technicians, citizen science is more cost-effective and can operate on broader scales on an ongoing basis. This makes it well-suited for conducting "big-picture" ecology and long-term monitoring. As the world becomes increasingly urbanized (*e.g.* Montgomery 2008), the recreational activities promoted by citizen science are also becoming increasingly important in that they promote natural awareness (Evans *et al.* 2005), which lends itself to advocacy and wildlife protection.

To ensure the usefulness of citizen science – and to maintain its credibility (Figure 2.1) – we must aggressively pursue and mitigate its sources of error. In this thesis, I identified several novel sources of error that could bias data collected from ornithological citizen science projects such as the Atlas of the Breeding Birds of Ontario ('OBBA'; Bird Studies Canada *et al.* 2008), the North American Breeding Bird Survey ('BBS'; Peterjohn 1994) and the Audubon Christmas Bird Count ('CBC'; Dunn *et al.* 2005). In my data chapters, showed (i) how observer skill and overconfidence might lead to false-positive errors of rare and common species (Chapter 3); (ii) how older observers might detect fewer birds than younger ones, and furthermore, how this may have led to excessively-negative population trend estimates in existing data (Chapter 4); (iii) how solitary observers beginning their volunteer service increase their counts gradually over about 5 years, and then decrease their counts substantially as they age (Chapter 5), (iv) how groups of observers might artificially inflate species richness over many consecutive, annual surveys (Chapter 5), and finally, (v) how modern volunteer birdwatchers might be just as skilled as their historical counterparts, contrary to the conclusions of previous research (Chapter 6).

Based on these findings, I make several recommendations and also consider future research directions. In Chapter 3, I emphasize the importance of accounting for observer skill when modeling survey data from multi-observer populations. This is especially critical in ‘omnibus’ surveys where multiple species are surveyed by a single observer, because observers of different skill levels are differentially prone to detect rare species, and so make false-positive detection errors for rare and common species at different relative frequencies. I also show how self-assessments of observer skill levels can function as accurate measures of ability in mixed models predicting detection errors.

In Chapters 2 and 3, I also describe how survey designs requiring repeated observer visits to a survey site can be a robust methodological approach to correct for inter-observer differences such as variable skill levels. This approach has been used successfully in several modern European surveys (*e.g.* the Swiss Survey of Common Breeding Birds; Royle *et al.* 2007) and in several Canadian bird atlasses, and allows modelers to explicitly calculate detection probabilities alongside the probability that a species is actually present. I recommend adopting such repeat-observer survey design elements where feasible. Furthermore, I recommend that surveys collect measures of detection uncertainty (sensu Miller *et al.* 2011) – which I also show in Chapter 3 must not be based upon an observer’s subjective opinion – so that modelers can simultaneously account for false-positive and false-negative detections (*e.g.* Miller *et al.* 2011). These errors can be quite numerous in survey data (McClintock *et al.* 2010*a*; Campbell and Francis 2011, Chapter 3), and can mislead conservation and management efforts (McKelvey *et al.* 2008). Future studies should explore how new designs and statistical methods can remove the influences of false-positive errors and overconfidence in survey data – as well as their confounding factors such as observer skill and species rarity – from real datasets.

In Chapter 4, I call for improvements in data collection and modeling that will take into account the ages and detection abilities of participating observers and correct for changes in these variables over time. I show how generalized additive mixed models (‘GAMMs’; Wood 2006) can be efficiently used to accomplish this goal, for instance by incorporating correction factors similar to Figure 4.5 and the ISO standard hearing-loss curves (Figure 4.1; International Organization for Standardization 2000). Because not all species detections are strictly auditory, and accordingly, because species tend to differ in their conspicuousness (Stewart 1954), future studies should test for the relative importance of hearing, sight, and other components of the detection process, and how they vary with the age of the observer and with the behaviour of specific species.

Similar to the recommendations made in Chapter 4 that modelers account for the age of observers, in Chapter 5, I recommend that modelers account for observer or party years of service at a given survey site. This approach can help correct for the count declines related to aging in later years of BBS service, and also to control for learning effects on both the BBS and CBC. Future modeling research on BBS data could determine whether modeling years of service as a sole covariate to account for learning and aging effects is sufficiently comprehensive, or whether a hybrid approach which also involves corrections for late-term aging might be superior.

There are many recommendations that have been made in the past that might improve the suitability of CBC data for scientific analysis (Francis *et al.* 2004; Dunn *et al.* 2005). Given my findings in Chapter 5 showing an increase in CBC richness over time occurring as a function of party years-of-service, I advocate for two in particular that were featured in a recent, formal review of the survey (Francis *et al.* 2004). First, I support the recommendation to establish standardized sub-surveys with more-rigorous effort controls within a given count circle. The more objective, consistent methods and transects in these sub-surveys would make them less-likely to see artificially-increased species counts with increasing years of party service. Second, I support the recommendation to better educate participants to appreciate the consequences of poor effort controls, and so to motivate them to be more consistent in their survey strategies outside of these controlled transects. However, because the recreational goals of the CBC, including its competitive nature, are firmly-rooted in its history (Bonta 2010), it will always be difficult to strike a balance between survey consistency and surveyor enjoyment (see Chapter 2), which directly affects participation levels (Dunn *et al.* 2005).

As a complement or an alternative to major design changes to the CBC, I also argue that GAMMs should be used to estimate party years-of-service effects if the number of consistent survey years is known. Future research could validate the years-of-service effects I observed in the CBC by considering data from external surveys that have similar methodological flexibility and competitiveness, for instance eBird (Sullivan *et al.* 2009).

In light of insights derived from Chapters 4 and 5 indicating that BBS observer abilities tend to decline with observer age, in Chapter 6, I show how the expected counts among newly-indoctrinated cohorts of BBS observers may be more-consistent over time than previous research has suspected. This chapter serves as a teaching point emphasizing the importance of recognizing that long-term patterns of within-observer error are found in BBS data. With new statistical methods currently being adopted for the analysis of BBS counts (Sauer and Link 2011), this thesis is a conveniently-timed resource that I hope will contribute to this process.

A major methodological focus of this thesis was on the use of GAMMs, a relatively new statistical technique which I show to be a useful alternative to parametric strategies, including hierarchical Bayesian approaches. GAMMs incorporate nonlinear smooth functions within broader parent functions that can optimally represent irregular patterns alongside other sources of variation. This ability makes them well-suited to the noisy data structure of wildlife survey data, and I show here how they can recognize and display nonlinear changes in observer behaviour that otherwise can bias population trend estimates in significant ways. This research is thus important not only for its demonstration of new and significant sources of observer error, but also for showing how GAMMs can help account for these errors. I hope that the merits of GAMMs will be considered by the modeling community as new statistical strategies are developed.

In this thesis, I worked with large datasets collected, in most cases, over several decades in hundreds of different survey locations. Consequently, the greatest vulnerability of this research is its reliance on consistent data collection conditions. Any systematic biases that are not measured or accounted for during data screening and analysis, for instance inconsistent noise levels over time (Griffith *et al.* 2010), long-term habitat changes (Keller and Scallan 1999), and undetected protocol changes (Gibbons *et al.* 2007) could have led to inaccuracies in the data. That said, all three datasets used (BBS, CBC, OBBA) are subject to rigorous scientific oversight which helps to identify and/or remove major intrinsic sources of error. However, as this thesis and past research (*e.g.* Kendall *et al.* 1996) demonstrate, ensuring the quality of citizen science data is an evolving and continuous process.

The results of my thesis on bird survey errors may also be applicable to citizen science projects surveying other taxa. For instance, anuran surveys are similar to bird surveys in that they rely heavily on auditory identifications, and also often use volunteers to achieve broad spatial coverage (*e.g.* de Solla *et al.* 2005). Some botanical monitoring projects also use large numbers of volunteers with different skill levels and motivations (Fitzpatrick *et al.* 2009), as do invertebrate studies (*e.g.* Dennis *et al.* 2006), and reptile monitoring programmes (Kéry *et al.* 2009). My research shows the importance of accounting for inter- and intra-observer variation in order to make accurate estimates of ecological states, and it offers methods to achieve this. It thus helps to make more-efficient use of research and/or monitoring funding, and of the well-intentioned and selfless efforts of thousands of dedicated volunteers.

Bibliography

Ahrends, A., C. Rahbek, M. T. Bulling, N. D. Burgess, P. J. Platts, J. C. Lovett, V. W. Kindemba, N. Owen, A. N. Sallu, A. R. Marshall, B. E. Mhoro, E. Fanning and R. Marchant. 2011. Conservation and the botanist effect. *Biological Conservation* **144**(1): 131–140. doi:10.1016/j.biocon.2010.08.008.

Alldredge, M. W., K. Pacifici, T. R. Simons and K. H. Pollock. 2008. A novel field evaluation of the effectiveness of distance and independent observer sampling to estimate aural avian detection probabilities. *Journal of Applied Ecology* **45**(5): 1349–1356. doi:10.1111/j.1365-2664.2008.01517.x.

Alldredge, M. W., K. H. Pollock and T. R. Simons. 2006. Estimating detection probabilities from multiple-observer point counts. *Auk* **123**(4): 1172–1182.

Alldredge, M. W., T. R. Simons and K. H. Pollock. 2007*a*. Factors affecting aural detections of songbirds. *Ecological Applications* **17**(3): 948–955.

Alldredge, M. W., T. R. Simons and K. H. Pollock. 2007*b*. A field evaluation of distance measurement error in auditory avian point count surveys. *Journal of Wildlife Management* **71**(8): 2759–2766. doi:10.2193/2006-161.

Balph, D. F. and M. H. Balph. 1983. On the psychology of watching birds: The problem of observer-expectancy bias. *Auk* **100**(3): 755–757.

Barker, R. J. and J. R. Sauer. 1992. Modeling population change from time series data. *In Wildlife 2001: populations*, eds. D. R. McCullough and R. Barrett. New York: Elsevier.

Barnosky, A. D., N. Matzke, S. Tomiya, G. O. U. Wogan, B. Swartz, T. B. Quental, C. Marshall, J. L. McGuire, E. L. Lindsey, K. C. Maguire, B. Mersey and E. A. Ferrer. 2011. Has the Earth's sixth mass extinction already arrived? *Nature* **471**(7336): 51–57. doi:10.1038/nature09678.

Bart, J. 1985. Causes of Recording Errors in Singing Bird Surveys. *Wilson Bulletin* **97**(2): 161–172.

Bart, J., K. P. Burnham, E. H. Dunn, C. M. Francis and C. J. Ralph. 2004*a*. Goals and strategies for estimating trends in landbird abundance. *Journal of Wildlife Management* **68**(3): 611–626.

Bart, J., B. Collins and R. I. G. Morrison. 2003. Estimating population trends with a linear model. *Condor* **105**(2): 367–372.

Bart, J., B. Collins and R. I. G. Morrison. 2004*b*. Estimating trends with a linear model: Reply to Sauer et al. *Condor* **106**(2): 440–443.

Bart, J. and J. D. Schoultz. 1984. Reliability of singing bird surveys - Changes in observer efficiency with avian density. *Auk* **101**(2): 307–318.

Bates, D. and M. Maechler. 2010. *lme4: Linear mixed-effects models using S4 classes.* URL http://CRAN.R-project.org/package=lme4.

Battersby, J. E. and J. J. D. Greenwood. 2004. Monitoring terrestrial mammals in the UK: past, present and future, using lessons from the bird world. *Mammal Review* **34**(1): 3–29.

Bell, S., M. Marzano, J. Cent, H. Kobierska, D. Podjed, D. Vandzinskaite, H. Reinert, A. Armaitiene, M. Grodziska-Jurczak and R. Muri. 2008. What counts? Volunteers and their organisations in the recording and monitoring of biodiversity. *Biodiversity and Conservation* **17**(14): 3443–3454. doi:10.1007/s10531-008-9357-9.

Betts, M. G., D. Mitchell, A. W. Diamond and J. Bety. 2007. Uneven rates of landscape change as a source of bias in roadside wildlife surveys. *Journal of Wildlife Management* **71**(7): 2266–2273.

Bevier, L. R., A. F. Poole and W. Moskoff. 2005. The Birds of North America: Veery (*Catharus fuscescens*)[online]. doi:10.2173/bna.142.

Bird Studies Canada, Environment Canada's Canadian Wildlife Service, Ontario Nature, Ontario Field Ornithologists and Ontario Ministry of Natural Resources. 2008. Atlas of the Breeding Birds of Ontario Database [online]. URL http://www.naturecounts.ca/.

Blaustein, A. R., D. B. Wake and W. P. Sousa. 1994. Amphibian declines - Judging stability, persistence, and susceptibility of populations to local and global extinctions. *Conservation Biology* **8**(1): 60–71.

Bonardi, A., R. Manenti, A. Corbetta, V. Ferri, D. Fiacchini, G. Giovine, S. Macchi, E. Romanazzi, C. Soccini, L. Bottoni, E. Padoa-Schioppa and G. F. Ficetola. 2011. Usefulness of volunteer data to measure the large scale decline of ”common” toad populations. *Biological Conservation* **144**(9): 2328–2334. doi:10.1016/j.biocon.2011.06.011.

Bonney, R., C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg and J. Shirk. 2009. Citizen science: A developing tool for expanding science knowledge and scientific literacy. *Bioscience* **59**(11): 977–984. doi:10.1525/bio.2009.59.11.9.

Bonta, M. 2010. Ornithophilia: Thoughts on geography in birding. *Geographical Review* **100**(2): 139–151.

Brand, A. R. 1938. Vibration frequencies of passerine bird song. *Auk* **55**(2): 263–268.

Braschler, B., K. Mahood, N. Karenyi, K. J. Gaston and S. L. Chown. 2010. Realizing a synergy between research and education: how participation in ant monitoring helps raise biodiversity awareness in a resource-poor country. *Journal of Insect Conservation* **14**(1): 19–30. doi:10.1007/s10841-009-9221-6.

Burnham, K. P. and D. R. Anderson. 2002. *Model selection and multimodal inference: a practical information-theoretic approach*. New York: Springer.

Butcher, G. S., M. R. Fuller, L. S. McAllister and P. H. Geissler. 1990. An evaluation of the Christmas Bird Count for monitoring population trends of selected species. *Wildlife Society Bulletin* **18**(2): 129–134.

Bystrak, D. 1981. The North American Breeding Bird Survey. *Studies in Avian Biology* **6**: 34–41.

Callahan, J. S., A. L. Brownlee, M. D. Brtek and H. L. Tosi. 2003. Examining the unique effects of multiple motivational sources on task performance. *Journal of Applied Social Psychology* **33**(12): 2515–2535.

Campbell, M. and C. Francis. 2011. Using stereo microphones to evaluate observer variation in North American Breeding Bird Survey (BBS) point counts. *Auk* **128**(2): 303–312. doi:10.1525/auk.2011.10005.

Carver, E. 2009. Birding in the United States: A demographic and economic analysis. Addendum to the 2006 Survey of Fishing, Hunting, and Wildlife-Associated Recreation. Tech. Rep. 2006-4, Arlington, VA.

Child, M. F., G. S. Cumming and T. Amano. 2009. Assessing the broad-scale impact of agriculturally transformed and protected area landscapes on avian taxonomic and functional richness. *Biological Conservation* **142**(11): 2593–2601. doi:10.1016/j.biocon.2009.06.007.

Clarke, E. D., L. B. Spear, M. L. Mccracken, F. F. C. Marques, D. L. Borchers, S. T. Buckland and D. G. Ainley. 2003. Validating the use of generalized additive models and at-sea surveys to estimate size and temporal trends of seabird populations. *Journal of Applied Ecology* **40**(2): 278–292.

Cohn, J. P. 2008. Citizen science: Can volunteers do real research? *Bioscience* **58**(3): 192–197. doi:10.1641/b580303.

Cooper, C. B., W. M. Hochachka and A. A. Dhondt. 2007. Contrasting natural experiments confirm competition between house finches and house sparrows. *Ecology* **88**(4): 864–870.

Croskerry, P. 2002. Achieving quality in clinical decision making: Cognitive strategies and detection of bias. *Academic Emergency Medicine* **9**(11): 1184–1204.

Cruickshanks, K. J., D. M. Nondahl, T. S. Tweed, T. L. Wiley, B. E. K. Klein, R. Klein, R. Chappell, D. S. Dalton and S. D. Nash. 2010. Education, occupation, noise exposure history and the 10-yr cumulative incidence of hearing impairment in older adults. *Hearing Research* **264**(1–2): 3–9. doi:10.1016/j.heares.2009.10.008.

Cyr, A. 1981. Limitation and variability in hearing ability in censusing birds. *Studies in Avian Biology* **6**: 327–333.

Dawson, D. K. and M. G. Efford. 2009. Bird population density estimated from acoustic signals. *Journal of Applied Ecology* **46**(6): 1201–1209. doi:10.1111/j.1365-2664.2009.01731.x.

de Solla, S. R., L. J. Shirose, K. J. Fernie, G. C. Barrett, C. S. Brousseau and C. A. Bishop. 2005. Effect of sampling effort and species detectability on volunteer based anuran monitoring programs. *Biological Conservation* **121**(4): 585–594. doi:10.1016/j.biocon.2004.06.018.

Dennis, R. L. H., T. G. Shreeve, N. J. B. Isaac, D. B. Roy, P. B. Hardy, R. Fox and J. Asher. 2006. The effects of visual apparency on bias in butterfly recording and monitoring. *Biological Conservation* **128**(4): 486–492. doi:10.1016/j.biocon.2005.10.015.

Dickinson, J. L., B. Zuckerberg and D. N. Bonter. 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics* **41**(1): 149–172. doi:10.1146/annurev-ecolsys-102209-144636.

Diefenbach, D. R., D. W. Brauning and J. A. Mattice. 2003. Variability in grassland bird counts related to observer differences and species detection rates. *Auk* **120**(4): 1168–1179.

Diefenbach, D. R., M. R. Marshall, J. A. Mattice and D. W. Brauning. 2007. Incorporating availability for detection in estimates of bird abundance. *Auk* **124**(1): 96–106.

Donald, P. F. and R. J. Fuller. 1998. Ornithological atlas data: a review of uses and limitations. *Bird Study* **45**(2): 129–145.

Downes, C. M. 2004. Results of the 2004 questionnaire for Canadian participants in the Breeding Bird Survey. URL http://www.cws-scf.ec.gc.ca/nwrc-cnrf/default.asp?lang=En&n=929AA800-1.

Duarte, C. M., J. Cebrian and N. Marba. 1992. Uncertainty of detecting sea-change. *Nature* **356**(6366): 190–190.

Dunn, E. H. 1995. Bias in Christmas Bird Counts for species that visit feeders. *Wilson Bulletin* **107**(1): 122–130.

Dunn, E. H. 2002. Using decline in bird populations to identify needs for conservation action. *Conservation Biology* **16**(6): 1632–1637.

Dunn, E. H., C. M. Francis, P. J. Blancher, S. R. Drennan, M. A. Howe, D. Lepage, C. S. Robbins, K. V. Rosenberg, J. R. Sauer and A. G. Smith. 2005. Enhancing the scientific value of the Christmas Bird Count. *Auk* **122**(1): 338–346.

Dunn, E. H., J. Larivee and A. Cyr. 1996. Can checklist programs be used to monitor populations of birds recorded during the migration season? *Wilson Bulletin* **108**(3): 540–549.

Dunn, E. H., J. Larivee and A. Cyr. 2001. Site-specific observation in the breeding season improves the ability of checklist data to track population trends. *Journal of Field Ornithology* **72**(4): 547–555.

Eglington, S. M., S. E. Davis, A. C. Joys, D. E. Chamberlain and D. G. Noble. 2010. The effect of observer experience on English Breeding Bird Survey population trends. *Bird Study* **57**(2): 129–141.

Ellis, R. and C. Waterton. 2005. Caught between the cartographic and the ethnographic imagination: the whereabouts of amateurs, professionals, and nature in knowing biodiversity. *Environment and Planning D: Society and Space* **23**(5): 673–693. doi:10.1068/d353t.

Elphick, C. S. 2008. How you count counts: the importance of methods research in applied ecology. *Journal of Applied Ecology* **45**(5): 1313–1320. doi:10.1111/j.1365-2664.2008.01545.x.

Emlen, J. T. and M. J. DeJong. 1992. Counting birds - the problem of variable hearing abilities. *Journal of Field Ornithology* **63**(1): 26–31.

Etterson, M. A., G. J. Niemi and N. P. Danz. 2009. Estimating the effects of detection heterogeneity and overdispersion on trends estimated from avian point counts. *Ecological Applications* **19**(8): 2049–2066. doi:10.1890/08-1317.1.

Evans, C., E. Abrams, R. Reitsma, K. Roux, L. Salmonsen and P. P. Marra. 2005. The Neighborhood Nestwatch program: Participant outcomes of a citizen-science ecological research project. *Conservation Biology* **19**(3): 589–594.

Faaborg, J., R. T. Holmes, A. D. Anders, K. L. Bildstein, K. M. Dugger, S. A. Gauthreaux, P. Heglund, K. A. Hobson, A. E. Jahn, D. H. Johnson, S. C. Latta, D. J. Levey, P. P. Marra, C. L. Merkord, E. Nol, S. I. Rothstein, T. W. Sherry, T. S. Sillett, F. R. Thompson and N. Warnock. 2010. Conserving migratory land birds in the New World: Do we know enough? *Ecology* **20**(2): 398–418. doi:10.1890/09-0397.1.

Faanes, C. A. and D. Bystrak. 1981. The role of observer bias in the North American Breeding Bird Survey. *Studies in Avian Biology* **6**: 353–359.

Farmer, R. G., M. L. Leonard and A. G. Horn. 2012. Observer effects and avian call count survey quality: rare-species biases and overconfidence. *Auk* **129**(1): 76–86. doi:10.1525/auk.2012.11129.

Farnsworth, G. L., K. H. Pollock, J. D. Nichols, T. R. Simons, J. E. Hines and J. R. Sauer. 2002. A removal model for estimating detection probabilities from point-count surveys. *Auk* **119**(2): 414–425.

Ferrer, X., L. M. Carrascal, O. Gordo and J. Pino. 2006. Bias in avian sampling effort due to human preferences: An analysis with Catalonian birds (1900-2002). *Ardeola* **53**(2): 213–227.

Fewster, R. M., S. T. Buckland, G. M. Siriwardena, S. R. Baillie and J. D. Wilson. 2000. Analysis of population trends for farmland birds using generalized additive models. *Ecology* **81**(7): 1970–1984.

Field, S. A., P. J. O'Connor, A. J. Tyre and H. P. Possingham. 2007. Making monitoring meaningful. *Austral Ecology* **32**(5): 485–491. doi:10.1111/j.1442-9993.2007.01715.x.

Field, S. A., A. J. Tyre, N. Jonzen, J. R. Rhodes and H. P. Possingham. 2004. Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters* **7**(8): 669–675. doi:10.1111/j.1461-0248.2004.00625.x.

Fitzpatrick, M. C., E. L. Preisser, A. M. Ellison and J. S. Elkinton. 2009. Observer bias and the detection of low-density populations. *Ecological Applications* **19**(7): 1673–1679. doi:10.1890/09-0265.1.

Flemming, J. M., E. Cantoni, C. Field and I. McLaren. 2010. Extracting long-term patterns of population changes from sporadic counts of migrant birds. *Environmetrics* **21**(5): 482–492. doi:10.1002/env.998.

Francis, C. M., J. Bart, E. H. Dunn, K. P. Burnham and C. J. Ralph. 2005. Enhancing the value of the Breeding Bird Survey: Reply to Sauer et al. (2005). *Journal of Wildlife Management* **69**(4): 1327–1332.

Francis, C. M., P. J. Blancher and R. D. Phoenix. 2009. Bird monitoring programs in Ontario: What have we got and what do we need? *The Forestry Chronicle* **85**(2): 202–217.

Francis, C. M., E. H. Dunn, P. J. Blancher, S. R. Drennan, M. A. Howe, D. Lepage, C. S. Robbins, K. V. Rosenberg, J. R. Sauer and K. G. Smith. 2004. Improving the Christmas Bird Count: Report of a review panel. *American Birds* **58**: 34–43.

Freeman, S. N., D. G. Noble, S. E. Newson and S. R. Baillie. 2007. Modelling population changes using data from different surveys: the common birds census and the breeding bird survey. *Bird Study* **54**: 61–72.

Gates, G. A. and J. H. Mills. 2005. Presbycusis. *Lancet* **366**(9491): 1111–1120.

Gates, G. A., P. Schmid, S. G. Kujawa, B. H. Nam and R. D'Agostino. 2000. Longitudinal threshold changes in older men with audiometric notches. *Hearing Research* **141**(1-2): 220–228.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models (Comment on an Article by Browne and Draper). *Bayesian Analysis* **1**(3): 515–533.

Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* **27**(15): 2865–2873. doi:10.1002/sim.3107.

Gelman, A. and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, United Kingdom: Cambridge University Press.

Gelman, A., A. Jakulin, M. G. Pittau and Y. S. Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* **2**(4): 1360–1383. doi:10.1214/08-AOAS191.

Gelman, A., Y. Su, M. Yajima, J. Hill, M. G. Pittau, J. Kerman and T. Zheng. 2010. *arm: Data analysis using regression and multilevel/hierarchical models*. URL http://CRAN.R-project.org/package=arm.

Genet, K. S. and L. G. Sargent. 2003. Evaluation of methods and data quality from a volunteer-based amphibian call survey. *Wildlife Society Bulletin* **31**(3): 703–714.

Gibbons, D. W., P. F. Donald, H. G. Bauer, L. Fornasari and I. K. Dawson. 2007. Mapping avian distributions: the evolution of bird atlases. *Bird Study* **54**: 324–334.

Goffredo, S., F. Pensa, P. Neri, A. Orlandi, M. Scola Gagliardi, A. Velardi, C. Piccinetti and F. Zaccanti. 2010. Unite research with what citizens do for fun: "Recreational monitoring" of marine biodiversity. *Ecological Applications* **20**(8): 2170–2187. doi:10.1890/09-1546.

Greenberg, R. and S. Droege. 1999. On the decline of the rusty blackbird and the use of ornithological literature to document long-term population trends. *Conservation Biology* **13**(3): 553–559.

Greenwood, J. 2007. Citizens, science and bird conservation. *Journal of Ornithology* **148**(S1): 77–124. doi:10.1007/s10336-007-0239-9.

Griffith, E. H., J. R. Sauer and J. A. Royle. 2010. Traffic effects on bird counts on North American Breeding Bird Survey routes. *Auk* **127**(2): 387–393. doi:10.1525/auk.2009.09056.

Gu, W. D. and R. K. Swihart. 2004. Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biological Conservation* **116**(2): 195–203. doi:10.1016/s0006-3207(03)00190-3.

Hanowski, J. A. M. and G. J. Niemi. 1995. A comparison of on- and off-road bird counts: Do you need to go off road to count birds accurately? *Journal of Field Ornithology* **66**(4): 469–483.

Haselmayer, J. and J. S. Quinn. 2000. A comparison of point counts and sound recording as bird survey methods in Amazonian southeast Peru. *Condor* **102**(4): 887–893.

Hewson, C., A. Amar, J. A. Lindsell, R. M. Thewlis, S. Butler, K. Smith and R. J. Fuller. 2007. Recent changes in bird populations in British broadleaved woodland. *Ibis* **149**(S2): 14–28. doi:10.1111/j.1474-919X.2007.00745.x.

Hull, J. M., A. M. Fish, J. J. Keane, S. R. Mori, B. N. Sacks and A. C. Hull. 2010. Estimation of species identification error: Implications for raptor migration counts and trend estimation. *Journal of Wildlife Management* **74**(6): 1326–1334. doi:10.2193/2009-255.

International Organization for Standardization. 2000. *Acoustics- Statistical distribution of hearing thresholds as a function of age (ISO 7029:2000)*. Geneva, Switzerland.

James, F. C., C. E. McCullogh and D. A. Wiedenfeld. 1996. New approaches to the analysis of population trends in land birds. *Ecology* **77**(1): 13–27.

Jiguet, F. 2009. Method learning caused a first-time observer effect in a newly started breeding bird survey. *Bird Study* **56**(2): 253–258. doi:10.1080/00063650902791991.

Johnson, D. H. 2008. In defense of indices: The case of bird surveys. *Journal of Wildlife Management* **72**(4): 857–868.

Joseph, L. N., S. A. Field, C. Wilcox and H. P. Possingham. 2006. Presence-absence versus abundance data for monitoring threatened species. *Conservation Biology* **20**(6): 1679–1687.

Joseph, L. N. and H. P. Possingham. 2008. Grid-based monitoring methods for detecting population declines: Sensitivity to spatial scale and consequences of scale correction. *Biological Conservation* **141**(7): 1868–1875.

Julliard, R., J. Clavel, V. Devictor, F. Jiguet and D. Couvet. 2006. Spatial segregation of specialists and generalists in bird communities. *Ecology Letters* **9**(11): 1237–1244.

Kampstra, P. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software* **28**(1): 1–9.

Keller, C. M. E. and J. T. Scallan. 1999. Potential roadside biases due to habitat changes along breeding bird survey routes. *Condor* **101**(1): 50–57.

Kendall, W. L., B. G. Peterjohn and J. R. Sauer. 1996. First-time observer effects in the North American Breeding Bird Survey. *Auk* **113**(4): 823–829.

Kepler, C. B. and J. M. Scott. 1981. Reducing bird count variability by training observers. *Studies in Avian Biology* **6**: 366–371.

Kéry, M., M. Dorazio, Robert, L. Soldaat, A. van Strien, A. Zuiderwijk and J. A. Royle. 2009. Trend estimation in populations with imperfect detection. *Journal of Applied Ecology* **46**(6): 1163–1172. doi:10.1111/j.1365-2664.2009.01724.x.

Kéry, M. and J. A. Royle. 2010. Hierarchical modelling and estimation of abundance and population trends in metapopulation designs. *Journal of Animal Ecology* **79**(2): 453–461. doi:10.1111/j.1365-2656.2009.01632.x.

Kéry, M., J. A. Royle, H. Schmid, M. Schaub, B. Volet, G. Hafliger and N. Zbinden. 2010. Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology* **24**(5): 1388–1397. doi:10.1111/j.1523-1739.2010.01479.x.

Kéry, M. and H. Schmid. 2004. Monitoring programs need to take into account imperfect species detectability. *Basic and Applied Ecology* **5**(1): 65–73.

Kéry, M. and H. Schmid. 2006. Estimating species richness: calibrating a large avian monitoring programme. *Journal of Applied Ecology* **43**(1): 101–110.

Klimkiewicz, M. K. and C. S. Robbins. 1978. Standard abbreviations for common names of birds. *North American Bird Bander* **3**(1): 16–25.

Kremen, C., K. S. Ullman and R. W. Thorp. 2011. Evaluating the quality of citizen-scientist data on pollinator communities. *Conservation Biology* **25**(3): 607–617. doi:10.1111/j.1523-1739.2011.01657.x.

La Rouche, G. P. 2001. Birding in the United States: A demographic and economic analysis. Addendum to the 2001 Survey of Fishing, Hunting, and Wildlife-Associated Recreation. Tech. Rep. 2001-1, Washington, DC.

La Sorte, F. A., T. M. Lee, H. Wilman and W. Jetz. 2009. Disparities between observed and predicted impacts of climate change on winter bird assemblages. *Proceedings of the Royal Society B: Biological Sciences* **276**(1670): 3167–3174. doi:10.1098/rspb.2009.0162.

La Sorte, F. A. and M. L. McKinney. 2007. Compositional changes over space and time along an occurrence-abundance continuum: Anthropogenic homogenization of the North American avifauna. *Journal of Biogeography* **34**(12): 2159–2167.

LaDeau, S. L., A. M. Kilpatrick and P. P. Marra. 2007. West Nile virus emergence and large-scale declines of North American bird populations. *Nature* **447**(7145): 710–U13.

Lane, K. A., J. Kang and M. R. Banaji. 2007. Implicit social cognition and law. *Annual Review of Law and Social Science* **3**: 427–451. doi:10.1146/annurev.lawsocsci.3.081806.112748.

Larrick, R. P., K. A. Burson and J. B. Soll. 2007. Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes* **102**(1): 76–94. doi:10.1016/j.obhdp.2006.10.002.

Lepage, D. and C. M. Francis. 2002. Do feeder counts reliably indicate bird population changes? 21 years of winter bird counts in Ontario, Canada. *Condor* **104**(2): 255–270.

Lindenmayer, D. B. and G. E. Likens. 2010. Improving ecological monitoring. *Trends in Ecology & Evolution* **25**(4): 200–201. doi:10.1016/j.tree.2009.11.006.

Lindenmayer, D. B., G. E. Likens, C. J. Krebs and R. J. Hobbs. 2010. Improved probability of detection of ecological "surprises". *Proceedings of the National Academy of Sciences of the United States of America* **107**(51): 21,957–21,962. doi:10.1073/pnas.1015696107.

Link, W. A. and J. R. Sauer. 1997*a*. Estimation of population trajectories from count data. *Biometrics* **53**(2): 488–497.

Link, W. A. and J. R. Sauer. 1997*b*. New approaches to the analysis of population trends in land birds: Comment. *Ecology* **78**(8): 2632–2634.

Link, W. A. and J. R. Sauer. 1998. Estimating population change from count data: Application to the North American Breeding Bird Survey. *Ecological Applications* **8**(2): 258–268.

Link, W. A. and J. R. Sauer. 1999. Controlling for varying effort in count surveys - An analysis of Christmas Bird Count data. *Journal of Agricultural, Biological, and Environmental Statistics* **4**(2): 116–125.

Link, W. A. and J. R. Sauer. 2002. A hierarchical analysis of population change with application to Cerulean Warblers. *Ecology* **83**(10): 2832–2840.

Link, W. A. and J. R. Sauer. 2007. Seasonal components of avian population change: joint analysis of two large-scale monitoring programs. *Ecology* **88**(1): 49–55.

Link, W. A., J. R. Sauer and D. K. Niven. 2006. A hierarchical model for regional analysis of population change using Christmas Bird Count data, with application to the American Black Duck. *Condor* **108**(1): 13–24.

Link, W. A., J. R. Sauer and D. K. Niven. 2008. Combining breeding bird survey and Christmas Bird Count data to evaluate seasonal components of population change in northern bobwhite. *Journal of Wildlife Management* **72**(1): 44–51. doi:10.2193/2007-299.

Lotz, A. and C. R. Allen. 2007. Observer bias in anuran call surveys. *Journal of Wildlife Management* **71**(2): 675–679. doi:10.2193/2005-759.

Lunn, D., D. Spiegelhalter, A. Thomas and N. Best. 2009. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* **28**(25): 3049–3067. doi:10.1002/sim.3680.

MacKenzie, D. I. 2005. What are the issues with presence-absence data for wildlife managers? *Journal of Wildlife Management* **69**(3): 849–860.

MacKenzie, D. I., J. D. Nichols, M. E. Seamans and R. J. Gutirrez. 2009. Modeling species occurrence dynamics with multiple states and imperfect detection. *Ecology* **90**(3): 823–835.

MacKenzie, D. I., J. D. Nichols, N. Sutton, K. Kawanishi and L. L. Bailey. 2005. Improving inferences in population studies of rare species that are detected imperfectly. *Ecology* **86**(5): 1101–1113.

Magurran, A. E., S. R. Baillie, S. T. Buckland, J. M. Dick, D. A. Elston, E. M. Scott, R. I. Smith, P. J. Somerfield and A. D. Watt. 2010. Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution* **25**(10): 574–582.

Maritimes Breeding Bird Atlas. 2006–2010. URL http://www.mba-aom.ca/.

Maritimes Butterfly Atlas. 2011. URL http://www.accdc.com/butterflyatlas.html.

Marsden, S. J. 1999. Estimation of parrot and hornbill densities using a point count distance sampling method. *Ibis* **141**(3): 377–390.

Mayfield. 1979. Amateur in ornithology. *Auk* **96**(1): 168–171.

Mayfield, H. 1966. Hearing loss and bird song. *Living Bird* **5**: 167–175.

McClintock, B. T., L. L. Bailey, K. H. Pollock and T. R. Simons. 2010*a*. Experimental investigation of observation error in anuran call surveys. *Journal of Wildlife Management* **74**(8): 1882–1893. doi:10.2193%2F2009-321.

McClintock, B. T., L. L. Bailey, K. H. Pollock and T. R. Simons. 2010*b*. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology* **91**(8): 2446–2454. doi:10.1890/09-1287.1.

McDonald-Madden, E., P. W. Baxter, R. A. Fuller, T. G. Martin, E. T. Game, J. Montambault and H. P. Possingham. 2010. Monitoring does not always count. *Trends in Ecology & Evolution* **25**(10): 547–550.

McKelvey, K., K. Aubry and M. Schwartz. 2008. Using anecdotal occurrence data for rare or elusive species: the illusion of reality and a call for evidentiary standards. *Bioscience* **58**(6): 549 – 555.

McLaren, A. A. and M. D. Cadman. 1999. Can novice volunteers provide credible data for bird surveys requiring song identification? *Journal of Field Ornithology* **70**(4): 481–490.

Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology* **92**(7): 1422–1428. doi:10.1890/10-1396.1.

Miller, D. T. and W. Turnbull. 1986. Expectancies and interpersonal processes. *Annual Review of Psychology* **37**: 233–256.

Montgomery, M. R. 2008. The urban transformation of the developing world. *Science* **319**(5864): 761–764. doi:10.1126/science.1153012.

Moore, D. A. and P. J. Healy. 2008. The trouble with overconfidence. *Psychological Review* **115**(2): 502–517. doi:10.1037/0033-295x.115.2.502.

Morris, J. C. and D. Q. McManus. 1991. The neurology of aging - Normal versus pathological change. *Geriatrics* **46**(8): 47–54.

Newson, S. E., K. L. Evans, D. G. Noble, J. J. D. Greenwood and K. J. Gaston. 2008. Use of distance sampling to improve estimates of national population sizes for common and widespread breeding birds in the UK. *Journal of Applied Ecology* **45**(5): 1330–1338. doi:10.1111/j.1365-2664.2008.01480.x.

Nichols, J. D., J. E. Hines, J. R. Sauer, F. W. Fallon, J. E. Fallon and P. J. Heglund. 2000. A double-observer approach for estimating detection probability and abundance from point counts. *Auk* **117**(2): 393–408.

Nichols, J. D., L. Thomas and P. B. Conn. 2009. Inferences about landbird abundance from count data: Recent advances and future directions. *In Environmental and Ecological Statistics*, eds. D. L. Thomson, E. G. Cooch and M. J. Conroy, vol. 3, pp. 201–235. Springer US. doi:10.1007/978-0-387-78151-8_9.

Nichols, J. D. and B. K. Williams. 2006. Monitoring for conservation. *Trends in Ecology & Evolution* **21**(12): 668–673. doi:10.1016/j.tree.2006.08.007.

Nisbet, E. 2007. Earth monitoring: Cinderella science. *Nature* **450**(6): 789–790. doi:10.1038/450789a.

Nondahl, D. A., X. Y. Shi, K. J. Cruickshanks, D. S. Dalton, T. S. Tweed, T. L. Wiley and L. L. Carmichael. 2009. Notched audiograms and noise exposure history in older adults. *Ear and Hearing* **30**(6): 696–703.

North American Amphibian Monitoring Program. 2011. URL http://www.pwrc.usgs.gov/naamp/.

Osei-Lah, V. and L. H. Yeoh. 2010. High frequency audiometric notch: An outpatient clinic survey. *International Journal of Audiology* **49**(2): 95–98. doi:10.3109/14992020903300423.

Pacifici, K., T. R. Simons and K. H. Pollock. 2008. Effects of vegetation and background noise on the detection process in auditory avian point-count surveys. *Auk* **125**(3): 600–607. doi:10.1525/auk.2008.07078.

Pautasso, M. and M. L. McKinney. 2007. The botanist effect revisited: Plant species richness, county area, and human population size in the United States. *Conservation Biology* **21**(5): 1333–1340. doi:10.1111/j.1523-1739.2007.00760.x.

Pearson, J. D., C. H. Morrell, S. Gordonsalant, L. J. Brant, E. J. Metter, L. L. Klein and J. L. Fozard. 1995. Gender differences in a longitudinal-study of age-associated hearing loss. *Journal of the Acoustical Society of America* **97**(2): 1196–1205.

Pellet, J. and B. R. Schmidt. 2005. Monitoring distributions using call surveys: Estimating site occupancy, detection probabilities and inferring absence. *Biological Conservation* **123**(1): 27–35.

Peterjohn, B. G. 1994. The North American Breeding Bird Survey. *Birding* **26**: 386–398.

Peterjohn, B. G. 2001. Some considerations on the use of ecological models to predict species' geographic distributions. *Condor* **103**(3): 661–663.

Peters, D. P. 2010. Accessible ecology: synthesis of the long, deep, and broad. *Trends in Ecology & Evolution* **25**(10): 592–601.

Peterson, A. P. 1995. Erroneous party-hour data and a proposed method of correcting observer effort in Christmas Bird Counts. *Journal of Field Ornithology* **66**(3): 385–390.

Pierce, B. A. and K. J. Gutzwiller. 2004. Auditory sampling of frogs: Detection efficiency in relation to survey duration. *Journal of Herpetology* **38**(4): 495–500.

Platt, J. R. 1964. Strong inference. *Science* **146**(3642): 347–353. doi:10.1126%2Fscience.146.3642.347.

Pollock, J. F. 2006. Detecting population declines over large areas with presence-absence, time-to-encounter, and count survey methods. *Conservation Biology* **20**(3): 882–892.

Preston, F. W. 1958. Analysis of the Audubon Christmas Counts in terms of the lognormal curve. *Ecology* **39**(4): 620–624.

Purcell, K. L., S. R. Mori and M. K. Chase. 2005. Design considerations for examining trends in avian abundance using point counts: Examples from oak woodlands. *Condor* **107**(2): 305–320.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. URL http://www.R-project.org.

Raitt, R. J. 1981. Chairman's introductory remarks: Observer variability. *Studies in Avian Biology* **6**: 326.

Ramsey, F. L. and J. M. Scott. 1981. Tests of hearing ability. *Studies in Avian Biology* **6**: 341–345.

Rempel, R. S., K. A. Hobson, G. Holborn, S. L. Van Wilgenburg and J. Elliott. 2005. Bioacoustic monitoring of forest songbirds: Interpreter variability and effects of configuration and digital processing methods in the laboratory. *Journal of Field Ornithology* **76**(1): 1–11.

Riddle, J. D., S. J. Stanislav, K. H. Pollock, C. E. Moorman and F. S. Perkins. 2010. Separating Components of the Detection Process With Combined Methods: An Example With Northern Bobwhite. *Journal of Wildlife Management* **74**(6): 1319–1325. doi:10.2193/2009-220.

Riffell, S. K. and B. D. Riffell. 2002. Can observer clothing color affect estimates of richness and abundance? An experiment with point counts. *Journal of Field Ornithology* **73**(4): 351–359.

Robbins, C. S. and R. W. Stallcup. 1981. Problems in separating species with similar habits and vocalizations. *Studies in Avian Biology* **6**: 360–365.

Rosenstock, S. S., D. R. Anderson, K. M. Giesen, T. Leukering and M. F. Carter. 2002. Landbird counting techniques: Current practices and an alternative. *Auk* **119**(1): 46–53.

Royle, J. A. and R. M. Dorazio. 2009. *Hierarchical modeling and inference in ecology: The analysis of data from populations, metapopulations and communities.* Academic Press (Elsevier).

Royle, J. A. and M. Kéry. 2007. A Bayesian state-space formulation of dynamic occupancy models. *Ecology* **88**(7): 1813–1823.

Royle, J. A., M. Kéry, R. Gautier and H. Schmid. 2007. Hierarchical spatial models of abundance and occurrence from imperfect survey data. *Ecological Monographs* **77**(3): 465–481.

Royle, J. A. and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology* **87**(4): 835–841.

Royle, J. A., J. D. Nichols and M. Kéry. 2005. Modelling occurrence and abundance of species when detection is imperfect. *Oikos* **110**(2): 353–359.

Ryder, T. B., R. Reitsma, B. Evans and P. P. Marra. 2010. Quantifying avian nest survival along an urbanization gradient using citizen- and scientist-generated data. *Ecology* **20**(2): 419–426. doi:10.1890/09-0040.1.

Sauer, J. R., J. E. Fallon and R. Johnson. 2003. Use of North American Breeding Bird Survey data to estimate population change for bird conservation regions. *Journal of Wildlife Management* **67**(2): 372–389.

Sauer, J. R. and W. A. Link. 2011. Analysis of the North American Breeding Bird Survey using hierarchical models. *Auk* **128**(1): 87–98. doi:10.1525/auk.2010.09220.

Sauer, J. R., W. A. Link, W. L. Kendall and D. D. Dolton. 2010. Comparative analysis of Mourning Dove population change in North America. *Journal of Wildlife Management* **74**(5): 1059–1069.

Sauer, J. R., W. A. Link, W. L. Kendall, J. R. Kelley and D. K. Niven. 2008. A hierarchical model for estimating change in American woodcock populations. *Journal of Wildlife Management* **72**(1): 204–214.

Sauer, J. R., W. A. Link, J. D. Nichols and J. A. Royle. 2005. Using the North American Breeding Bird Survey as a tool for conservation: A critique of Bart et al. (2004). *Journal of Wildlife Management* **69**(4): 1321–1326.

Sauer, J. R., W. A. Link and J. A. Royle. 2004. Estimating population trends with a linear model: Technical comments. *Condor* **106**(2): 435–440.

Sauer, J. R., G. W. Pendleton and B. G. Peterjohn. 1996. Evaluating causes of population change in North American insectivorous songbirds. *Conservation Biology* **10**(2): 465–478.

Sauer, J. R., B. G. Peterjohn and W. A. Link. 1994. Observer differences in the North American Breeding Bird Survey. *Auk* **111**(1): 50–62.

Schmeller, D. S., P. Henry, R. Julliard, B. Gruber, J. Clobert, F. Dziock, S. Lengyel, P. Nowicki, E. Deri, E. Budrys, T. Kull, K. Tali, B. Bauch, J. Settele, C. van Swaay, A. Kobler, V. Babij, E. Papastergiadou and K. Henle. 2009. Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology* **23**(2): 307–316. doi:10.1111/j.1523-1739.2008.01125.x.

Shirose, L. J., C. A. Bishop, D. M. Green, C. J. MacDonald, R. J. Brooks and N. J. Helferty. 1997. Validation tests of an amphibian call count survey technique in Ontario, Canada. *Herpetologica* **53**(3): 312–320.

Silvertown, J. 2009. A new dawn for citizen science. *Trends in Ecology & Evolution* **24**(9): 467–471. doi:10.1016/j.tree.2009.03.017.

Simons, T. R., M. W. Alldredge, K. H. Pollock and J. M. Wettroth. 2007. Experimental analysis of the auditory detection process on avian point counts. *Auk* **124**(3): 986–999.

Snall, T., O. Kindvall, J. Nilsson and T. Part. 2010. Evaluating citizen-based presence data for bird monitoring. *Biological Conservation* **144**(2): 804–810. doi:10.1016/j.biocon.2010.11.010.

Stewart, P. A. 1954. The value of Christmas Bird Counts. *Wilson Bulletin* **66**(3): 184–195.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink and S. Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* **142**(10): 2282–2292.

Thogmartin, W. E., B. R. Gray, M. Gallagher, N. Young, J. J. Rohweder and M. G. Knutson. 2007. Power to detect trend in short-term time series of bird abundance. *Condor* **109**(4): 943–948.

Thomas, L. 1996. Monitoring long-term population change: Why are there so many analysis methods? *Ecology* **77**(1): 49–58.

Thomas, L. and K. Martin. 1996. The importance of analysis method for breeding bird survey population trend estimates. *Conservation Biology* **10**(2): 479–490.

Toppila, E., I. Pyykk and J. Starck. 2001. Age and noise-induced hearing loss. *Scandinavian Audiology* **30**(4): 236–244. doi:10.1080/01050390152704751.

Trumbull, D. J., R. Bonney, D. Bascom and A. Cabral. 2000. Thinking scientifically during participation in a citizen-science project. *Science Education* **84**(2): 265–275. doi:10.1002/(SICI)1098-237X(200003)84:2⟨265::AID-SCE7⟩3.0.CO;2-5.

U.S. North American Bird Conservation Initiative Committee. 2010. The State of the Birds 2010 Report on Climate Change. URL http://www.stateofthebirds.org/State%20of%20the%20Birds%202011.pdf.

Van Eyken, E., G. Van Camp and L. Van Laer. 2007. The complexity of age-related hearing impairment: Contributing environmental and genetic factors. *Audiology and Neuro-Otology* **12**(6): 345–358. doi:10.1159/000106478.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*. New York: Springer. URL http://www.stats.ox.ac.uk/pub/MASS4.

Ward-Paige, C. A., C. Mora, H. K. Lotze, C. Pattengill-Semmens, L. McClenachan, E. Arias-Castro and R. A. Myers. 2010. Large-scale absence of sharks on reefs in the greater-Caribbean: A footprint of human pressures. *PLoS ONE* **5**(8): e11,968. doi:10.1371/journal.pone.0011968.

Weeks Jr., H. P. 1994. The Birds of North America: Eastern Phoebe (*Sayornis phoebe*) [online]. doi:10.2173/bna.94.

Wiedner, D. and P. Kerlinger. 1990. Economics of birding: A national survey of active birders. *American Birds* **44**(2): 209–213.

Wiersma, Y. F. 2010. Birding 2.0: Citizen science and effective monitoring in the Web 2.0 world. *Avian Conservation and Ecology* **5**(2): 13.

Wiley, T. L., R. Chappell, L. Carmichael, D. A. Nondahl and K. J. Cruickshanks. 2008. Changes in hearing thresholds over 10 years in older adults. *Journal of the American Academy of Audiology* **19**(4): 281–292. doi:10.3766/jaaa.19.4.2.

Wilson, S., S. LaDeau, A. Tottrup and P. Marra. 2011. Range-wide effects of breeding and non-breeding season climate on the abundance of a Neotropical migrant songbird. *Ecology* **92**(9): 1789–1798. doi:10.1890/10-1757.1.

Wood, S. N. 2006. *Generalized Additive Models: An Introduction with R.* New York: Chapman and Hall.

Wood, S. N. 2011. *gamm4: Generalized additive mixed models using mgcv and lme4.* URL http://CRAN.R-project.org/package=gamm4.

Yoccoz, N. G., J. D. Nichols and T. Boulinier. 2001. Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution* **16**(8): 446–453.

Ziolkowski Jr., D. and K. L. Pardieck. 2006. Volunteers drive the North American Breeding Bird Survey [poster presentation]. *In North American Ornithological Conference, Veracruz, Mexico.* URL http://www.pwrc.usgs.gov/bbs/bbsnews/MeetingProducts/Volunteer-poster-NAOC(final).pdf.

Zuur, A. F., E. N. Ieno, N. J. Walker, A. A. Saveliev and G. M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R.* New York: Springer.