8-2014

Applied Statistics in Environmental Monitoring: Case Studies and Analysis for the Michigan Bald Eagle Biosentinel Program

Katherine Leith
Clemson University, leith.kate@gmail.com

APPLIED STATISTICS IN ENVIRONMENTAL MONITORING: CASE STUDIES AND ANALYSIS FOR THE MICHIGAN BALD EAGLE BIOSENTINEL PROGRAM.

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Wildlife and Fisheries Biology

---

by
Katherine Ferne Leith
August 2014

Accepted by:
Dr. William W. Bowerman, Committee Co-Chair
Dr. William C. Bridges, Committee Co-Chair
Dr. Joseph D. Lanham
Dr. Webb M. Smathers

ABSTRACT

The bald eagle (*Haliaeetus leucocephalus*) is an extensively researched tertiary predator. Its life history and the impact of various stressors on its reproductive outcomes have been documented in many studies, and over many years. Furthermore, the bald eagle population recovery in Michigan has been closely monitored since the 1960s, as it has continued to recover from a contaminant-induced bottleneck. Because of its position at the top of the aquatic food web and the large body of ethological knowledge, the bald eagle has become a sentinel species for the Michigan aquatic ecosystem. In April 1999, the Michigan Department of Environmental Qualtity, Water Division, began monitoring environmentally persistent and toxic contaminants in bald eagles.

Continued monitoring of bald eagle population dynamics and contaminant levels in the environment are important to understanding the fate of sentinel species and ecosystems after exposure to environmental contaminants. It is therefore essential to develop sound methods of analysis to apply in reporting observations and in assessing trends based on these data. Specifically, this study assesses the Michigan Bald Eagle Biosentinel Program's (1) power to detect regionally elevated contaminant concentrations or assure remediation success; (2) various techniques for reporting central tendency in left-censored data using PCB and p,p'DDE contaminant concentrations; and the effects of model specification on inferential conclusions in regarding reproductive outcome as a function of site classification.

ACKNOWLEDGEMENTS

There have been so many sources of support over the time that I was working on this, and I am deeply grateful for all of them. Upon arriving in Clemson, Split Creek Farm gave me community, sustenance (both emotional and physical), and much joy. During my time living in South Carolina, class mates, who were always good for a laugh or a shoulder to cry on, provided priceless camaraderie. I'm also very lucky to have a brilliant family, all the members of which have provided perspective and empathy, as well as editorial support and countless margaritas.

I must thank my friends and colleagues at Ventana Medical Systems, Inc. Heather, Stephanie, Crystal, Scott, Bob, Szu-Yu, Chang, Bonnie, and many others have been great colleagues and great support. I am also thankful that they became accustomed to, and I dare say accepted, my antics with admirable ease! I especially thank Dr. James Ranger-Moore, who knew full well that this would take much of my energy and focus, but supported my finishing while in his employ anyway.

I thank my committee (Dr. Bowerman, Dr. Bridges, Dr. Smathers, and Dr. Lanham) for many things, not the least of which is patience. They have helped me grow, both academically and personally.

Dr. Michael Wierda, my husband, partner in crime, and dear friend, I thank for his unwavering support. While I cannot begin to guess at its location parameter, it is decidedly leptokurtic: a gift of inestimable value and low variance. He also likes my nerdy humor, no mean feat, which makes everything easier.

TABLE OF CONTENTS

LIST OF FIGURES

General Introduction.

Chapter 1.

Chapter 2.

Chapter 3.

LIST OF TABLES

Chapter 1.

Chapter 2.

Chapter 3.

PREFACE

This dissertation was written in journal style and organized into three chapters, each with an introduction, methods, results, and discussion. Each chapter is intended for publication and repetition in some sections (i.e. Introduction, Methods, Results, Discussion, and Literature Cited) may occur. The chapters are preceded by a General introduction and followed by overall Conclusions.

General Introduction

Bald Eagles and Biosentinels

The bald eagle *(Haliaeetus leucocephalus)* is notable both for its position in the ecosystem and in the public eye. It is a large bird of prey, considered to be piscivorous, but which also opportunistically forages on an array of avian, mammalian, and reptilian prey (Buehler, 2000). Territory size is difficult to estimate because methods of measurement are not consistent and nesting densities vary widely based on habitat and food supply (Buehler, 2000). Mean productivity has been estimated at 1.87 eggs per clutch and clutches usually range from one to three eggs (Stalmaster, 1987). Bald eagles are associated with aquatic habitats throughout North America including coastal areas, rivers, lakes, reservoirs, and forested shorelines (Buehler, 2000). Because it is a tertiary predator in these ecosystems, it is susceptible to biomagnification of a wide array of xenobiotics. Extensive research has been conducted on this high-profile raptor addressing life history characteristics and the influences of various stressors on reproduction.

The bald eagle was selected as a biosentinel species for monitoring contaminants in Michigan's surface waters for the following reasons:

1. As a top-level predator, the bald eagle has a significant reliance on the aquatic food web and feeds primarily on fish and waterbirds. Specific dietary preferences of bald eagles include species of northern pike *(Esox lucius)*, suckers *(Catostomus spp.)*, bullheads *(Ameiurus spp.)*, carp

(Hypopthalmichthys spp.), bowfin *(Amia calva)*, ducks *(*family: *Anatidae)*, gulls *(*family: *Laridae)*, and white-tailed deer *(Odocoileus virginianus)*, as winter carrion and road-kill.

2. Past monitoring has shown that eagles accumulate organic and inorganic environmental contaminants and those contaminants may be quantified in blood, feather, egg, and tissue samples.

3. There is a expanding population of bald eagles that provides sufficient sampling opportunities for a long-term monitoring program.

4. The large body size of nestling eagles (eaglets) allows monitoring to be conducted by sampling blood and sufficient sample volumes are available to attain low quantification levels.

5. Mature bald eagles display great fidelity to their nesting territory and often return to the same nest tree year after year. Some wintering eagles may move away from their nesting territories, however many reside within the state's waters throughout the year. Once nesting and breeding has been initiated in spring and the breeding pair has returned to a breeding area, they defend and hunt within their territory.

These attributes of bald eagle ecology in Michigan, in addition to the fact that our samples are from pre-fledged eagles, support the conclusion that contaminants found in nestling bald eagles will represent the uptake of available contaminants within a particular territory. For all of these reasons the bald eagle is an excellent biosentinel species.

Long Term Monitoring

The bald eagle endured several threats to its population through the 20th century. In the early 1900s, many eagles were shot and as the country grew, industrialized human encroachment on habitat became a limiting factor in their distribution. After World War II, the use of pesticides was fairly widespread, though the damage they caused to the ecosystem was not yet fully understood. Experimental evidence was published in the 1950's showing that the reproductive success of birds can be affected by steady intake of DDT (Dewitt, 1956, 1955; Genelly & Rudd, 1956). By the 1960's, both citizens and the scientific community had become aware of the precipitous drop in bald eagle numbers in the Great Lakes region. This large charismatic raptor, which had once maintained active breeding territories every 8 to 16 km along the coasts in Michigan, had been reduced to just 82 occupied territories in 1972 (Postupalsky, 1989). With the publication of Rachel Carson's *Silent Spring* in 1962, the decline in bird populations was presented to the general public as a consequence of pesticide use. In 1966, The Journal of Applied Ecology published a special supplemental issue entitled "*Pesticides in the Environment and Their Effect on Wildlife*," which included explorations of DDT residues in birds, the

effect of pesticides on Lake Michigan's ecosystem, and the importance of developing a pesticide monitoring program (Bernard, 1966; Hickey et al., 1966; Keith, 1966; Moore, 1966). Largely because of the public awareness created by *Silent Spring*, DDT use was declining even before it was banned in 1972 and the positive effects could be seen in Michigan's bald eagle population. Though many of the organochlorine compounds were banned in the early 1970s, they are extremely persistent in the environment (Grier, 1982). Monitoring of contaminant levels, both through time and at various spatial scales, provides important insight into the health of the Great Lakes Basin ecosystem.

There have been numerous studies on the detrimental effects of persistent chemicals on different avian species (Cope, 2004). By 1990, work had been published exploring the relationship between observed contaminant residues in bald eagle eggs and shell thinning and reproduction (Grubb et al., 1990; Wiemeyer et al., 1993, 1984). In 1993, a review of both ecological and toxicological factors regulating bald eagle productivity in the Great Lakes Basin highlighted the primary factors influencing bald eagle populations: habitat availability, contaminant concentration, and degree of human disturbance (Bowerman, 1993). It could be shown that bald eagles were limited by habitat availability and food abundance, as well as that in the presence of plentiful food, eagles would occupy suboptimal habitat (Hansen, 1987; Newton, 1979; Stalmaster, 1987). Concentrations of PCB 126 (3,3',4,4',5-pentachlorobiphenyl) found in bald eagles eggs were approximately 20 times higher than the lowest toxic concentration tested in American kestrels (*Falco sparverius*) and may be a factor in the decline of some eagle populations (Hoffman et al., 1996). Bald eagles were shown to have normal young

production when egg DDE concentrations were < 3.6 μg/g (wet weight). When egg concentrations were between 3.6 and 6.3 μg/g production was halved and production was halved again when egg concentrations were > 6.3 μg/g (Wiemeyer et al., 1993). It has now been well established that high contaminant concentrations are associated with low rates of productivity in bald eagles (Bowerman, 1993; Bowerman et al., 2003, 1995, 1994, 1993; Dykstra et al., 2001; Grubb et al., 1992, 1990; Grubb & King, 1991; Wiemeyer et al., 1984).

Careful observation by ornithologists revealed more complexity to the recovery trends. It appeared that nesting pairs along the Great Lakes coast were rebounding less successfully than inland populations (Postupalsky, 1985). This suspicion was later confirmed as studies of contaminant concentrations from collected eggs, and later from nestlings, showed higher contaminant concentrations in areas with a Great Lakes centered prey base when compared to inland areas (Best et al., 1994; Bowerman et al., 2003, 1995; Wiemeyer et al., 1993, 1984). This phenomenon set the stage for a source sink dynamic in the Great Lakes basin in which the less contaminated inland regions of the state supply sufficient young to keep the Great Lakes coastal populations growing in spite of reproductive productivity rates which still demonstrate impairment (Bowerman et al., 2003; Simon, 2013).

The State of Michigan has maintained a count of occupied bald eagle breeding areas and their reproductive outcomes that extends back to 1961. Terminology used in this dissertation regarding bald eagle productivity and territories follows that of Postupalsky

(1974) . These records, initially collected during bald eagle nestling banding efforts, document the growth in the number of active and successful breeding areas. As nest numbers grew, accurate nest and outcome assessments could no longer be conducted by visiting every nest within the state during the breeding period and in 1977 the U.S. Fish and Wildlife Service began flights for nest and nestling enumeration. Arial enumeration continues on a yearly basis, though it is now carried out through the Michigan Departments of Natural Resources and Environmental Quality. In spite of the fact that Michigan is home to roughly 600 occupied breeding areas each season, outcomes are still obtained for every known occupied breeding area and 100-200 nests are visited by ground crews, which serves to verify the flight based assessments of outcome. The longevity and thorough nature of the data collection effort in Michigan have made it an extremely powerful information source.

Nesting eagles are found along the shorelines and on islands of each of the four Great Lakes surrounding Michigan. Further, the distribution of breeding eagles across much of Michigan's interior provides monitoring coverage for many of the major river systems (Figure 1). Currently, active bald eagle breeding areas are well distributed across the Upper Peninsula and northern Lower Peninsula of Michigan. While the breeding areas there has also continued to increase as eagles either establish new establishment of breeding areas in southern Michigan took longer, the number of active breeding areas there has also continued to increase as eagles either establish new

breeding areas or re-occupy historical territories.

Figure 1. Locations of bald eagle breeding territories throughout the state of Michigan, 2009.

In April 1999, the Michigan Department of Environmental Qualtity (MDEQ), Water Division, began monitoring environmentally persistent and toxic contaminants in bald eagles. This study is part of the wildlife contaminant monitoring project component of the MDEQ's Nonpoint Source Environmental Monitoring Strategy (MDEQ, 2004). The November 1998 passage of the Clean Michigan Initiative-Clean Water Fund (CMI-CWF) bond proposal resulted in a substantial increase in annual funding for a statewide surface water quality monitoring program beginning in 1999. The CMI-CWF offers reliable funding for the monitoring of surface water quality over a period of approximately 15 years. This is important because one of the goals of the Strategy is to measure temporal and spatial trends in contaminant levels in Michigan's surface waters.

Annually a subset of the active territories is sampled and eagle plasma and feathers are analyzed for mercury, PCBs, and chlorinated pesticides including DDT. Also, efforts are being made to expand the analyte list to include emerging contaminants such as brominated or fluorinated compounds. Watersheds with eagle nests and successful reproduction are assessed once every five years consistent with the National Pollutant Discharge Elimination System NPDES five-year basin cycle (Figure 2). This sampling procedure is consistent with that of other monitoring projects conducted within the designated watersheds under the NPDES permitting process (MDEQ, 1997). Nests associated with the Great Lakes and connecting channels are sampled annually because of the uncertainty of nesting success from year to year. An annual report is prepared that describes spatial and temporal trends in productivity and contaminant levels. In accordance with one of the key principles of the CMI-CWF, the bald eagle monitoring

protocol was planned and conducted in partnership with outside organizations. In 1999, this partnership included Lake Superior State University and Clemson University, from 2000 to 2008 this partnership included Michigan State University and Clemson University, from 2009 to 2012 this was conducted solely by Clemson University, and since 2013 it has been conducted by the University of Maryland.

Complications of Data Management

Continued monitoring of contaminant levels in the environment is an important part of understanding the fate of ecosystems after contamination. It is therefore essential to develop sound methods of analysis to detect meaningful changes in contaminant levels.

As levels decrease over time, limitations in analytical equipment create a lower bound below which contaminant levels cannot be accurately reported. This results in datasets with observations below the detection limit (DL) that are reported only as 'non-detect' or '< DL' and no value is provided. This type of distribution is called 'left-censored', as the low-end observations that are unknown generally occur near the origin of the x-axis in figures.

Several options for the analysis of these datasets have been investigated leading to the conclusion that methods replacing all non-detects with a single value (substitution methods) are frequently inferior (Antweiler & Taylor, 2008; Baccarelli et al., 2005; Eastoe et al., 2006; Helsel, 2006, 2005b, 2005a; Liu et al., 1997; Needham et al., 2007; Singh & Nocerino, 2002).

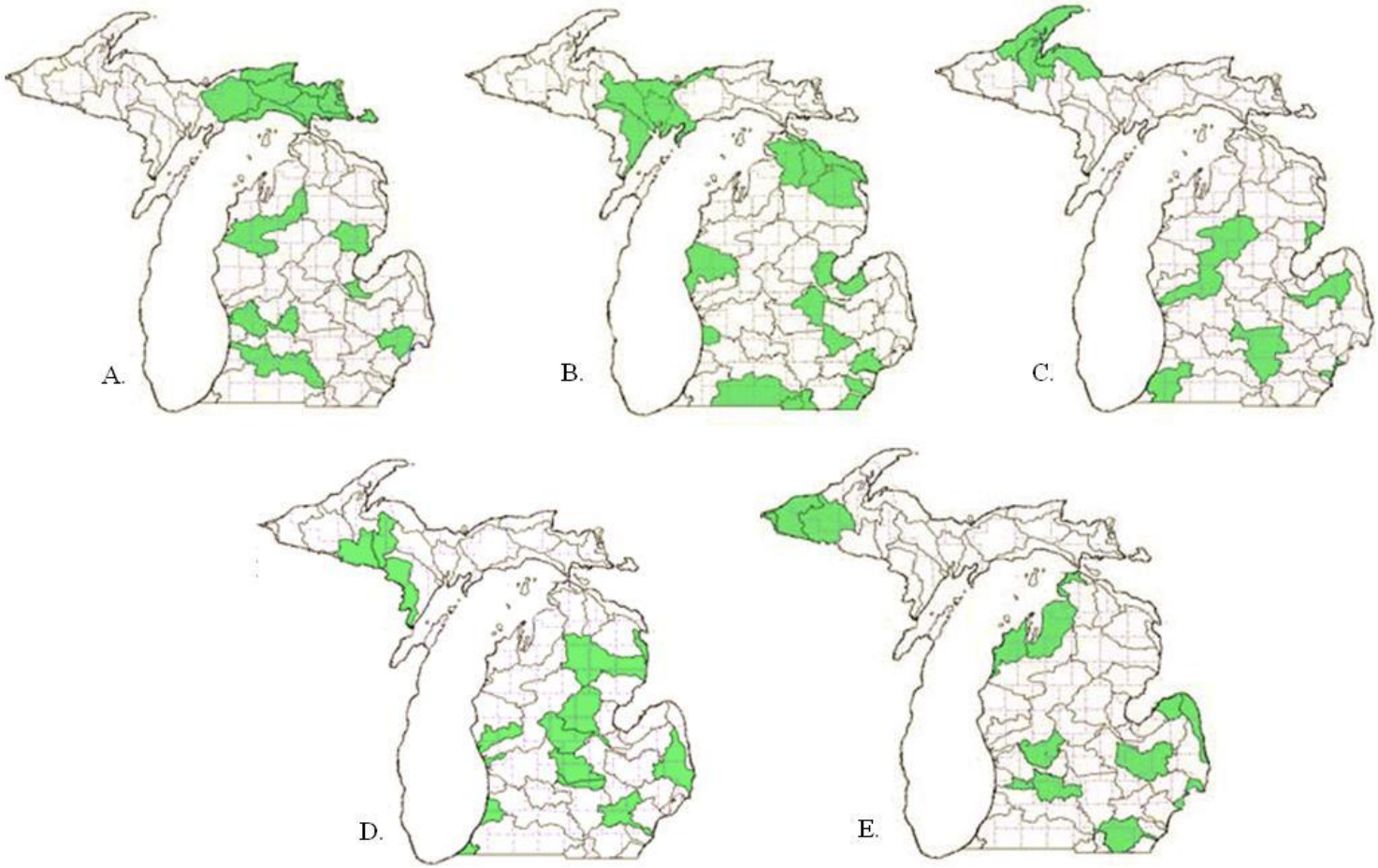

Figure 2. Michigan's watershed delineations and monitoring 'basin years'. A.) 1999, 2004 basin year watersheds (shaded); B.) 2000, 2005 basin year watersheds (shaded); C.) 2001, 2006 basin year watersheds (shaded); D.) 2002, 2007 basin year watersheds (shaded); and E.) 2003, 2008 basi n year watersheds (shaded).

Specifically, it has been shown that the bias caused by substitution increases dramatically as the percent of observations censored increases (Eastoe et al., 2006). In spite of this, various substitution methods continue to be used in research, frequently with little regard for the proportion of observations censored.

In addition to censored observations, another complication when analyzing environmental contaminant data is the possibility that datasets display a right skew. This distribution is common in environmental data and can frequently be accommodated by log-transformation. A final complexity is added by the fact that lognormal data are frequently summarized using the geometric mean, which is particularly sensitive to the choice of substitution value. Current statistical methods include tests of significant differences between regions at several geographic scales, and calculating descriptive statistics in the form of geometric means. Substitution is currently used in cases of non-detect, or left-censored, observations.

This choice of methods for addressing the non-detects does not affect testing of significant differences among regions. The monitoring program uses nonparametric Kruskal-Wallis and Wilcoxon tests which are rank-based and make no assumptions about distribution and so, are not sensitive to the problems of substitution (Helsel, 2005b). However, summary statistics are reported as geometric means, which are affected by the choice of substituted value. The desire to summarize data more accurately has fueled recent comparisons of proposed analytical alternatives to substitution or simple median reporting (Helsel, 2005a).

Complications of Reproductive Trend Assessment

It is common practice in environmental monitoring to use summary statistics, such as means, in modeling trends. The underlying data are often not normally distributed (for example: counts, presence/absence), but a normal distribution is used under the assumption that sample sizes will sufficiently normalize the summary data being modeled. For large samples this is likely the case, but what constitutes sufficiently 'large' may vary based on the underlying complexities of the data source. While the Central Limit Theorem suggests that as sample size approaches 30 distributions of statistics approach normality, monitoring data often consist of correlated (clustered) observations, such as repeated measurements made on the same site. This violates the assumption of independence of observations that is fundamental to parametric inferential analysis. When analyzing summary data, no adjustment can be made for this underlying correlation structure. Violating either or both of these assumptions of normality and independence can undermine the validity of significance tests.

Objectives

This dissertation is organized into a general introductory chapter, three chapters consisting of stand-alone manuscripts, and a summary chapter. The objectives of each chapter were as follows:

Chapter 1: Assessment of Michigan Bald Eagle Biosentinel Program's power to detect regionally elevated contaminant concentrations or assure remediation success.

The objectives were to:

(1) Assess the fit of lognormal distribution to the data;

(2) Determine a reasonable estimate for standard deviation based on existing observations;

(3) Estimate sample sizes necessary to produce analyze data with a power of 0.80, 0.85, 0.90, and 0.95 based on changes of 10%, 15%, 20%, and 25% in observed concentration; and

(4) Provide reference tables of recommended sample sizes based on those results.

Chapter 2: A comparison of techniques for assessing central tendency in left-censored data using PCB and p,p'DDE contaminant concentrations from Michigan's Bald Eagle Biosentinel Program.

The objectives were to:

(1) Assess the fit of lognormal distribution to the data;

(2) Compare and contrast the performance of the four methods of non-detect handling in terms of estimated geometric mean, comparison to the median, and standard error; and

(3) Make recommendations based on those results.

Chapter 3: Assessment of the effects of model specification on inferential conclusions in regarding reproductive outcome as a function of site classification.

The objectives were to:

(1) Determine appropriate levels of analysis for regional comparison of reproductive trends based on the needs and interests of the ongoing monitoring effort;

(2) Assess the fit of normal distribution to the summary statistics derived from raw data;

(3) Fit models to the raw data accounting for source distribution for reproductive outcomes, and correlated measures do to repeated sampling within sites; and

(4) Assess potential impact of model specification to inferences made regarding regional differences in trends for reproductive outcomes as compared to inferences that would result from such comparisons made based on summarized data.

Literature Cited

Antweiler, R. C., & Taylor, H. E. (2008). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. Environmental Science & Technology, 42(10), 3732-3738.

Baccarelli, A., Pfeiffer, R., Consonni, D., Pesatori, A. C., Bonzini, M., Patterson, D. G., Bertazzi, P. A., Landi, M. T. (2005). Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the Seveso chloracne study. Chemosphere, 60(7), 898-906.

Bernard, R. F. (1966). DDT residues in avian tissues. Journal of Applied Ecology, 193-198.

Best, D. A., Bowerman, W., Kubiak, T. J., Winterstein, S. R., Postupalsky, S., Shieldcastle, M., & Giesy, J. (1994). Reproductive impairment of bald eagles Haliaeetus leucocephalus along the Great Lakes shorelines of Michigan and Ohio. Raptor Conservation Today, 697-702.

Bowerman, W. W. (1993). Regulation of bald eagles (Haliaeetus leucocephalus) productivity in the Great Lakes Basin: And ecological and toxicological approach. Unpublished PhD dissertation, Michigan State University, East Lansing, MI, USA.

Bowerman, W. W., Best, D. A., Giesy, J. P., Shieldcastle, M. C., Meyer, M. W., Postupalsky, S., & Sikarskie, J. G. (2003). Associations between regional differences in polychlorinated biphenyls and dichlorodiphenyldichloroethylene in blood of nestling bald eagles and reproductive productivity. Environmental Toxicology and Chemistry, 22(2), 371-376.

Bowerman, W. W., Giesy, J. P., Best, D. A., & Kramer, V. J. (1995). A review of factors affecting productivity of bald eagles in the Great-Lakes region – Implications for recovery. Environmental Health Perspectives, 103, 51-59.

Bowerman, W. W., Grubb, T. G., Bath, A. J., Giesy, J. P., Dawson, G. A., & Ennis, R. K. (1993). Population composition and perching habitat of wintering bald eagles, *Haliaeetus-leucocephalus*, in Northcentral Michigan. Canadian Field-Naturalist, 107(3), 273-278.

Bowerman, W. W. I. V., Best, D. A., Giesy, J. P., Jr., Kubiak, T. J., Sikarskie, J. G., Meyburg, B. U., & Chancellor, R. D. (1994). The influence of environmental contaminants on bald eagle Haliaeetus leucocephalus populations in the Laurentian Great Lakes, North America. Raptor conservation today. 703-707.

Buehler, D. A. (2000). Bald Eagle: Haliaeetus leucocephalus. Birds of North America(506), 1-40.

Cope W. G., Leidy, R. B., and Hodgson, E. (2004). Classes of Toxicants: Use Classes. A textbook of modern toxicology. New Jersy: John Wiley & Sons: 49-73.

DeWitt, J. B. (1956). Pesticide Toxicity, Chronic Toxicity to Quail and Pheasants of Some Chlorinated Insecticides. Journal of Agricultural and Food Chemistry, 4(10), 863-866. doi: 10.1021/jf60068a004

Dewitt, J. B, Derby, J. V., Mangan, G. F. (1955). DDT vs. Wildlife. Relationships between quantities ingested, toxic effects and tissue storage. Journal of the American Pharmaceutical Association, 44(1), 22--24.

Dykstra, C. R., Meyer, M. W., Stromborg, K. L., Warnke, D. K., Bowerman, W. W., & Best, D. A. (2001). Association of low reproductive rates and high contaminant levels in bald eagles on Green Bay, Lake Michigan. Journal of Great Lakes Research, 27(2), 239-251.

Eastoe, E. F., Halsall, C. J., Heffernan, J. E., & Hung, H. (2006). A statistical comparison of survival and replacement analyses for the use of censored data in a contaminant air database: A case study from the Canadian Arctic. Atmospheric Environment, 40(34), 6528-6540.

Genelly, R. E., & Rudd, R. L. (1956). Effects of DDT, toxaphene, and dieldrin on pheasant reproduction. The Auk, 529-539.

Grier, J. W. (1982). Ban of DDT and subsequent recovery of Reproduction in bald eagles. Science, 218(4578), 1232-1235.

Grubb, T. G., Bowerman, W. W., Giesy, J. P., & Dawson, G. A. (1992). Responses of breeding bald eagles, *Haliaeetus-leucocephalus,* to human activities in Northcentral Michigan. Canadian Field-Naturalist, 106(4), 443-453.

Grubb, T. G., & King, R. M. (1991). Assessing human disturbance of breeding bald eagles with classification tree models. Journal of Wildlife Management, 55(3), 500-511.

Grubb, T. G., Wiemeyer, S. N., & Kiff, L. F. (1990). Eggshell thinning and contaminant levels in bald eagle eggs from Arizona, 1977 to 1985. Southwestern Naturalist, 35(3), 298-301.

Hansen, A. J. (1987). Regulation of bald eagle reproductive rates in Southeast Alaska. Ecology, 68(5), 1387-1392.

Helsel, D. R. (2005a). More than obvious: Better methods for interpreting nondetect data. Environmental Science & Technology, 39(20), 419A-423A.

Helsel, D. R. (2005b). Nondetects and Data Analysis: Statistics for Censored Environmental Data. Hoboken: John Wiley & Sons, Inc.

Helsel, D. R. (2006). Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. Chemosphere, 65(11), 2434-2439.

Hickey, J. J., Keith, J., & Coon, F. B. (1966). An exploration of pesticides in a Lake Michigan ecosystem. Journal of Applied Ecology, 141-154.

Hoffman, D. J., Melancon, M. J., Klein, P. N., Rice, C. P., Eisemann, J. D., Hines, R. K., Spann, J. W., Pendleton, G. W. (1996). Developmental toxicity of PCB 126 (3,3',4,4',5-pentachlorobiphenyl) in nestling American kestrels (Falco sparverius). Fundamental and Applied Toxicology, 34(2), 188-200.

Keith, J. A. (1966). Reproduction in a population of herring gulls (Larus argentatus) contaminated by DDT. Journal of Applied Ecology, 57-70.

Keith, J. O. (1966). Insecticide contaminations in wetland habitats and their effects on fish-eating birds. Journal of Applied Ecology, 71-85.

Liu, S., Lu, J.-C., Kolpin, D. W., & Meeker, W. Q. (1997). Analysis of Environmental Data with Censored Observations. Environmental Science & Technology, 31(12), 3358-3362.

MDEQ. (1997). A strategic environmental quality monitoring program for Michigan's surface waters. (Vol. January 1997). Michigan Department of Environmental Quality, Lansing, Michigan.

MDEQ. (2004). Nonpoint Source Environmental Monitoring Strategy. MDEQ Submittal to U.S. Environmental Protection Agency (Vol. September 2004). Michigan Department of Environmental Quality, Lansing, Michigan.

Moore, N. (1966). A pesticide monitoring system with special reference to the selection of indicator species. Journal of Applied Ecology, 261-269.

Needham, L. L., Naiman, D. Q., Patterson, D. G., & LaKind, J. S. (2007). Assigning concentration values for dioxin and furan congeners in human serum when measurements are below limits of detection: An observational approach. Chemosphere, 67(3), 439-447.

Newton, I. (1979). Population Ecology of Raptors. Vermillion, S.D.: Buteo Books.

Postupalsky, S. (1974). Raptor reproductive success: some problems with methods, criteria, and terminology. Paper presented at the Conf. Raptor Conservation Techniques.

Postupalsky, S. (1989). Eagle and Osprey Research in Michigan, 1985: Michigan Department of Natural Resources, Wildlife Division, Natural Heritage Program.

Simon, K.L. (2013). Bald Eagle (Haliaeetus Leucocephalus) Population Productivity and Source-Sink Dynamics in Michigan, 1961-2010. SETAC North America 34th Annual Meeting, Nashville, Tennessee.

Singh, A., & Nocerino, J. (2002). Robust estimation of mean and variance using environmental data sets with below detection limit observations. Chemometrics and Intelligent Laboratory Systems, 60(1-2), 69-86.

Stalmaster, M. V. (1987). The bald eagle. New York: Universe Books.

Wiemeyer, S. N., Bunck, C. M., & Stafford, C. J. (1993). Environmental contaminants in bald eagle eggs 1980-84 and further interpretations of relationships to productivity and shell thickness. Archives of Environmental Contamination and Toxicology, 24(2), 213-227.

Wiemeyer, S. N., Lamont, T. G., Bunck, C. M., Sindelar, C. R., Gramlich, F. J., Fraser, J. D., & Byrd, M. A. (1984). Organochlorine pesticide, polychlorobiphenyl, and mercury residues in bald eagle eggs—1969–79—and their relationships to shell thinning and reproduction. Archives of Environmental Contamination and Toxicology, 13(5), 529-549.

Chapter 1: Assessment of Michigan Bald Eagle Biosentinel Program's power to detect regionally elevated contaminant concentrations or assure remediation success.

Introduction

The bald eagle *(Haliaeetus leucocephalus)* is notable both for its position in the ecosystem and in the public eye. It is a large bird of prey, considered to be piscivorous, but which also opportunistically forages on an array of avian, mammalian, and reptilian prey (Buehler, 2000). Territory size is difficult to estimate because methods of measurement are not consistent and nesting densities vary widely based on habitat and food supply (Buehler, 2000). Bald eagles are associated with aquatic habitats throughout North America including coastal areas, rivers, lakes, reservoirs, and forested shorelines (Buehler, 2000). Because it is a tertiary predator in these ecosystems, it is susceptible to biomagnification of a wide array of xenobiotics.

The bald eagle has been shown to be an appropriate model to monitor ecosystem contaminant concentrations. Great Lakes nestling bald eagles receive prey items from within the adult's local breeding territory. Concentrations of Bioaccumulative Contaminants of Concern (BCC) in nestling eagle feathers and blood plasma reflect exposure to BCCs from the food items they receive. The eagle is therefore an appropriate indicator of ecosystem quality (Bowerman et al., 1998; Roe, 2004). In addition, the fact that our samples are from pre-fledged eagles, support the conclusion that contaminants found in nestling bald eagles will represent the uptake of available contaminants within a

particular territory. For all of these reasons, the bald eagle was selected as a biosentinel species for monitoring contaminants in Michigan's surface waters (MDEQ, 1997).

There have been numerous studies on the detrimental effects of persistent chemicals on different avian species (Cope, 2004). The relationship between observed contaminant residues in bald eagle eggs collected across the U.S. and shell thinning and reproduction was studied at the Patuxant Wildlife Research Center (Grubb et al., 1990; Wiemeyer et al., 1993; Wiemeyer et al., 1984). Concentrations of total PCBs found in bald eagles eggs were approximately 20 times higher than the lowest toxic concentration tested in American kestrels (*Falco sparverius*) for PCB 126 (3,3',4,4',5-pentachlorobiphenyl) and may be a factor in the decline of some eagle populations (Hoffman et al., 1996). Bald eagles were shown to have normal young productivity (defined as number of fledged young per occupied nest) when egg DDE concentrations were < 3.6 μg/g (wet weight). When egg concentrations were between 3.6 and 6.3 μg/g productivity was halved and productivity was halved again when egg concentrations were > 6.3 μg/g (Wiemeyer et al., 1993). It has now been well established that high contaminant concentrations are associated with low rates of productivity in bald eagles (Bowerman, 1993; Bowerman et al., 2003, 1995, 1994; Dykstra et al., 2001; Grubb et al., 1990; Wiemeyer et al., 1984).

There is also evidence for spatial differences in contaminant concentrations. Studies of contaminant concentrations from collected eggs, and later from nestlings, showed higher contaminant concentrations in areas with a Great Lakes centered prey base

when compared to inland areas (Bowerman et al., 2003, 1995; Giesy et al., 1995). This phenomenon set the stage for a source sink dynamic in the Great Lakes basin in which the less contaminated inland regions of the state supply sufficient young to keep the Great Lakes coastal populations growing in spite of reproductive productivity rates which still demonstrate impairment (Bowerman et al., 2003).

In April 1999, the Michigan Department of Environmental Qualtity (MDEQ), Water Division, began monitoring environmentally persistent and toxic contaminants in bald eagles. This study is part of the wildlife contaminant monitoring project component of the MDEQ's Nonpoint Source Environmental Monitoring Strategy (MDEQ, 1997). The November 1998 passage of the Clean Michigan Initiative-Clean Water Fund (CMI-CWF) bond proposal resulted in a substantial increase in annual funding for a statewide surface water quality monitoring program beginning in 1999. The CMI-CWF offers reliable funding for the monitoring of surface water quality over a period of approximately 15 years. This is important because one of the goals of the Strategy is to measure temporal and spatial trends in contaminant levels in Michigan's surface waters.

Nesting eagles are found along the shorelines and on islands of each of the four Great Lakes surrounding Michigan. Further, the distribution of breeding eagles across much of Michigan's interior provides monitoring coverage for many of the major river systems (Figure 1).

Figure 2. Locations of bald eagle breeding territories throughout the state of Michigan, 2009.

Annually a subset of the active territories is sampled and eagle plasma and feathers are analyzed for mercury, PCBs, and chlorinated pesticides including DDT. Watersheds with eagle nests and successful reproduction are assessed once every five years consistent with the National Pollutant Discharge Elimination System NPDES five-year basin cycle (Figure 2). In accordance with one of the key principles of the CMI-CWF, the bald eagle monitoring protocol was planned and conducted in partnership with outside organizations. In 1999, this partnership included Lake Superior State University and Clemson University, from 2000 to 2008 this partnership included Michigan State University and Clemson University, from 2009 to 2012 this was conducted solely by Clemson University, and since 2013 it has been conducted by the University of Maryland.

The five-year watershed monitoring cycle allows for only a portion of the watersheds within the state of Michigan to be sampled every year. A complete cycle of five years of sampling data should be representative of the concentrations of contaminants and of productivity and success rates for the entire state. This comprehensive data set will be useful for making human health and wildlife management recommendations and decisions.

The overall objective of this study was to use Bald Eagle Biosentinel Program (BEBP) data to determine the sample sizes that would be necessary to detect regionally elevated contaminant concentrations when compared to reference site concentrations.

Specifically, the objectives of this investigation were:

(1) Assess the fit of lognormal distribution to the data;

(2) Determine a reasonable estimate of standard deviation on existing log-transformed observations;

(3) Estimate sample sizes necessary to analyze data with a power of 0.80, 0.85, 0.90, and 0.95 based on changes of 10%, 15%, 20%, and 25% in log concentration for balanced and unbalanced experimental designs; and

(4) Provide reference tables of recommended sample sizes based on those results.

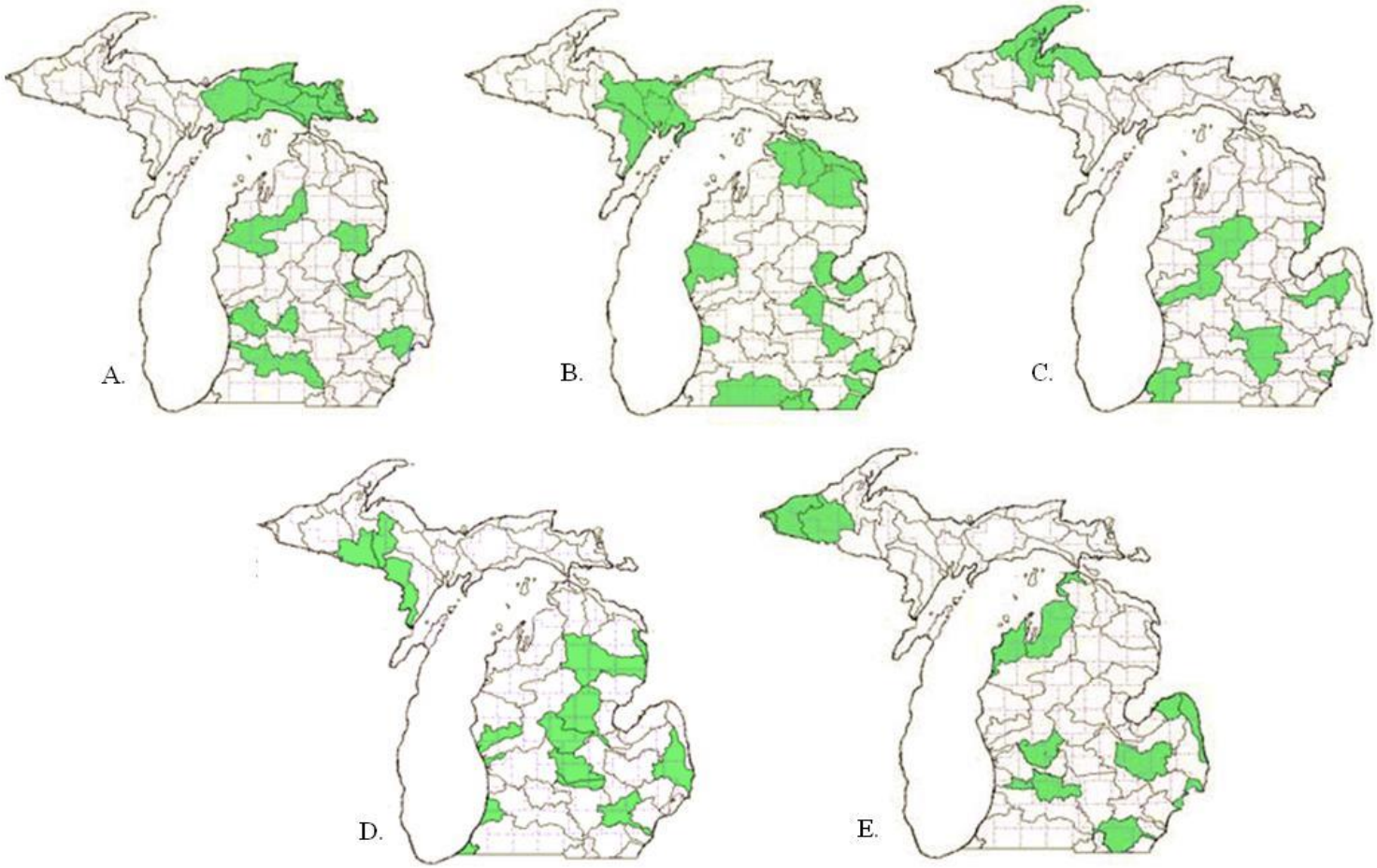

Figure 2. Michigan's watershed delineations and monitoring 'basin years'. A.) 1999, 2004 basin year watersheds (shaded); B.) 2000, 2005 basin year watersheds (shaded); C.) 2001, 2006 basin year watersheds (shaded); D.) 2002, 2007 basin year watersheds (shaded); and E.) 2003, 2008 basi n year watersheds (shaded).

Methods

Summary of Field and Analytical Methods

Nestling bald eagles were sampled from the Upper and Lower Peninsulas of Michigan and from the surrounding Michigan Islands. Blood was collected during normal banding activities from mid-May through late June from 1999-2003. Nestlings were between 5 and 10 weeks of age. Aseptic techniques were used to collect 10-13 cc of blood from the brachial vein with heparinized syringes fitted with 22 or 25 gauge needles. Morphometric measurements were used at this time to determine sex and age of the nestlings (Bartolotti 1984a, b). A total of 398 nestling eagles from 227 breeding areas were sampled and analyzed from 1999 to 2003. Samples of whole blood were transferred to heparinized vacuum tubes, stored on ice in coolers, and centrifuged within 48 hours of collection. Blood plasma was decanted, transferred to new heparinized vacuum tubes, sealed, and then frozen. All samples were shipped and stored at the U. S. Fish and Wildlife Service East Lansing Field Office until analysis at Clemson University (Roe, 2004).

All extractions and analyses were conducted according to procedures detailed in Clemson Institute of Environmental Toxicology (CIET 401-78-01) standard operating procedures. In brief, concentrations of organochlorine compounds were quantified by capillary gas chromatography with an electron capture detector using the United States Environmental Protection Agency approved methods. Chicken plasma was used for laboratory control samples in all analytical batches. All reported results were confirmed

by dual column analysis. Method validation studies were conducted on chicken plasma as a surrogate matrix to ensure that the data quality objectives of the Quality Assurance Project Plan (CIET 1996, 1999) were met (Wierda, 2009).

Statistical Methods

Total PCBs and total p,p'DDE log scale concentrations from 1999 through 2003 were utilized in estimating the parameters necessary for this sample size analysis. Concentrations of total PCBs and p,p'DDE less than the method detection limits were reported as non-detects and represented 6.41% and 10.26% of the 234 observations, respectively. Concentrations below the detection limit were set at ½ the detection limit based on the rates of censorship and the recommendations of Leith, et al. (2010).

The data analysis for this paper was generated using SAS® software, Version 9.1.2 of the SAS system for Windows (Copyright 2000-2004 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA).

Data were analyzed for significant departure from the lognormal distribution using the UNIVARIATE PROCEDURE with the options 'normal' and 'plot' activated (SAS Institute Inc., 2000-2004). This procedure produces several fit statistics. The Kolomogorov-Smirnov statistic was used to assess the assumption of fit to the lognormal distribution. The fit was classified as 'good' with p-values ≥ 0.05, 'marginal' with p-values between 0.01 and 0.05, and fit was rejected for p-values<0.01. In keeping with 'best practice' recommendations plots of the log-scale data were also inspected for

worrisome deviation from the assumed normal distribution (Farrell & Rogers-Stewart, 2006; Noughabi & Arghami, 2010; Razali & Wah, 2011; Romão et al., 2009; Seier, 2002).

Sample size analysis was performed using the POWER PROCEDURE (SAS Institute Inc., 2000-2004) to estimate sample sizes necessary to detect regionally elevated contaminant concentrations assuming the availability of a reference site, based on changes of 10%, 15%, 20%, and 25% in log concentration. Sample size estimates were generated for the above effect sizes with a power of 0.80, 0.85, 0.90, and 0.95 and for scenarios with unbalanced sampling ratios of experimental:reference sites from group weights of 1:1, 1:2, 1:3, and 1:4. This created a 4 by 4 by 4 matrix (64 different scenarios) of results for each of the two contaminants analyzed. Results were organized by contaminant and effect size. All estimates are based on an alpha of 0.05.

Results

Kolmogorov-Smirnov (KS) tests suggested that the lognormal distribution did not fit the PCB data and the p,p'DDE data equally well. The PCB distribution was marginal ($P=0.0206$) and the p,p'DDE distribution significantly differed from lognormal ($P<0.01$). This was likely due to the presence of six outliers in the upper tail of the p,p'DDE distribution. When these outliers were removed, the KS test resulted in no evidence for significant departure from lognormality ($P=0.1338$). Though log transformation did not successfully normalize the distribution, concentrations were positively skewed in a manner similar to log-normal distribution commonly seen in other contaminant research.

Furthermore, Figures 3 and 4, which show the log-scale data plotted against the normal distribution, do not suggest worrisome deviation from the assumed normal distribution, particularly in light of the known effect of substitution using ½ the detection limit for observations below the detection limit. For these reasons, and in keeping with environmental toxicology's convention of reporting geometric means, log scale concentrations were used in this analysis.

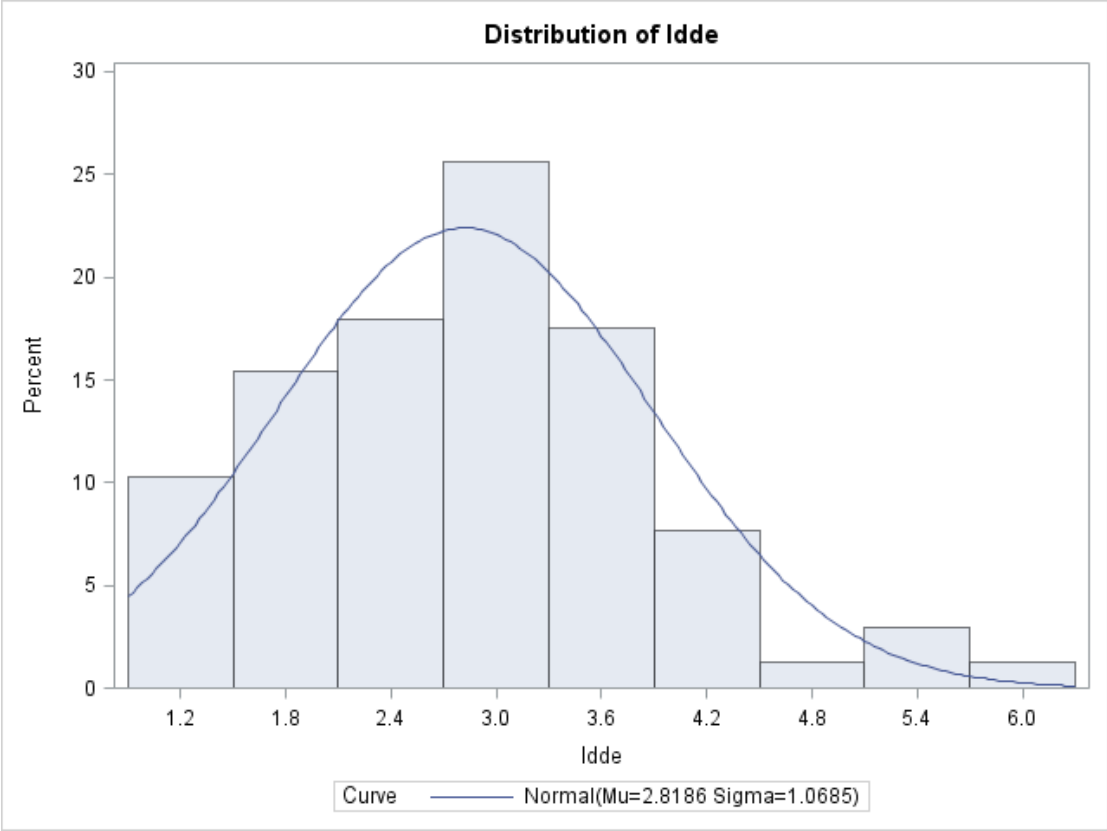

Figure 3. Distribution of log-scale p,p'DDE concentrations (ldde) and the normal distribution, for comparison.

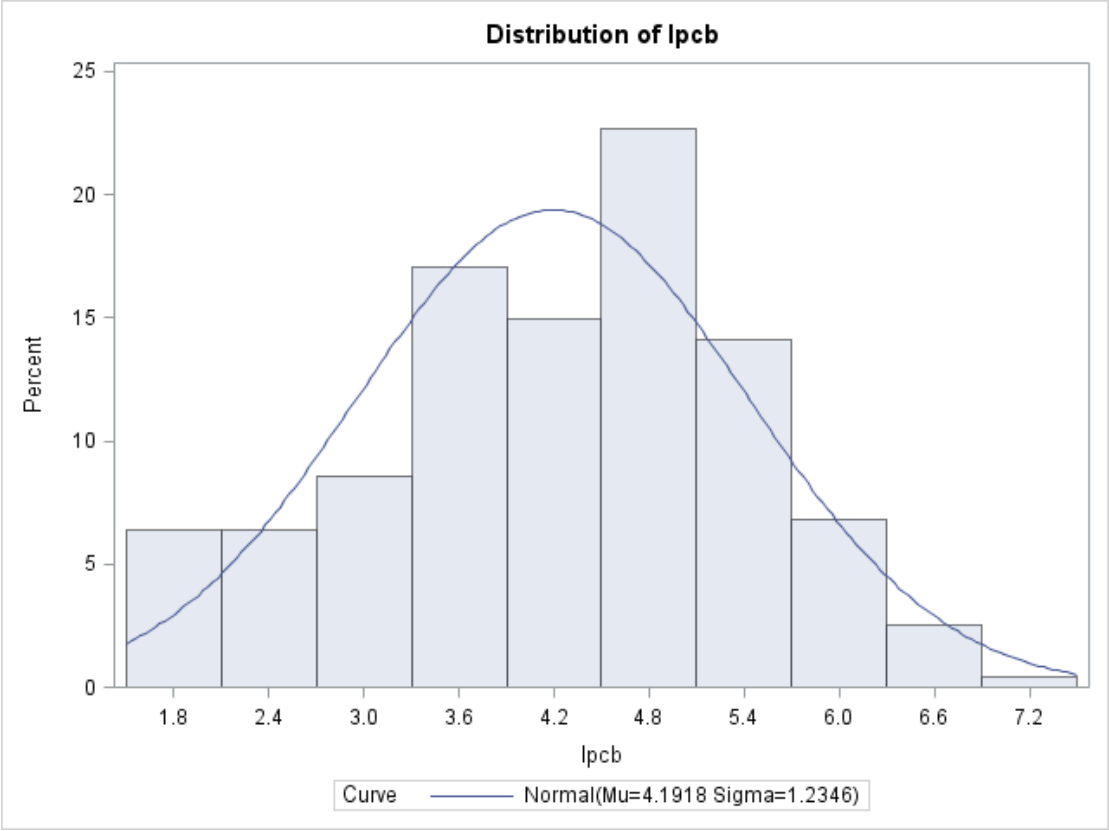

Figure 4. Distribution of log-scale PCB concentrations (lpcb) and the normal distribution, for comparison.

Standard deviation was calculated for the entire data set and for regional subsets since this investigation is intended to inform regional analysis. Estimates were largely in agreement for both contaminants and all regional scales, ranging from (log-scale) approximately 0.67 ppb at the low end to approximately 4.19 ppb at the high end. Since most of the estimates of standard deviation for regional subsets were close to the estimate based data from the entire state, the state-wide estimates of standard deviation were used.

The log-scale standard deviation for PCB concentrations was 1.23 ppb and the log-scale standard deviation estimate for p,p'DDE concentrations was 1.07.

For power analysis regarding p,p'DDE detection, results are summarized in Tables 1 through 4. Estimates of required sample sizes for p,p'DDE analysis are provided in Table 1 for balanced sampling structures. Table 2 provides estimates assuming twice the available number of reference sites to sites of suspected elevated contaminant concentration. Table 3 provides estimates assuming three times the available number of reference sites to sites of suspected elevated contaminant concentration. Table 4 provides estimates assuming four times the available number of reference sites to sites of suspected elevated contaminant concentration.

For power analysis regarding PCB detection, results are summarized in Tables 5 through 8. Estimates of required sample sizes for PCB analysis are provided in Table 5 for balanced sampling structures. Table 6 provides estimates assuming twice the available number of reference sites to sites of suspected elevated contaminant concentration. Table 7 provides estimates assuming three times the available number of reference sites to sites of suspected elevated contaminant concentration. Table 8 provides estimates assuming four times the available number of reference sites to sites of suspected elevated contaminant concentration.

Table 1. Sample sizes required for detecting differences in log PCB concentrations between a site of suspected exposure and a reference site provided a balanced sampling structure.

% Difference	Power	Total N	Exposed Site N
10	0.80	430	215
	0.85	492	246
	0.90	576	288
	0.95	710	355
15	0.80	198	99
	0.85	226	113
	0.90	264	132
	0.95	324	162
20	0.80	114	57
	0.85	130	65
	0.90	152	76
	0.95	186	93
25	0.80	74	37
	0.85	84	42
	0.90	98	49
	0.95	122	61

Table 2. Sample sizes required for detecting differences in log PCB concentrations between a site of suspected exposure and a reference site provided an unbalanced sampling structure with twice as many reference site samples.

% Difference	Power	Total N	Exposed Site N
10	0.80	483	161
	0.85	552	184
	0.90	648	216
	0.95	798	266
15	0.80	222	74
	0.85	255	85
	0.90	297	99
	0.95	366	122
20	0.80	129	43
	0.85	147	49
	0.90	171	57
	0.95	210	70
25	0.80	84	28
	0.85	96	32
	0.90	111	37
	0.95	135	45

Table 3. Sample sizes required for detecting differences in log PCB concentrations between a site of suspected exposure and a reference site provided an unbalanced sampling structure with three times as many reference site samples.

% Difference	Power	Total N	Exposed Site N
10	0.80	572	143
	0.85	656	164
	0.90	768	192
	0.95	948	237
15	0.80	264	66
	0.85	300	75
	0.90	352	88
	0.95	432	108
20	0.80	152	38
	0.85	172	43
	0.90	200	50
	0.95	248	62
25	0.80	100	25
	0.85	112	28
	0.90	132	33
	0.95	160	40

Table 4. Sample sizes required for detecting differences in log PCB concentrations between a site of suspected exposure and a reference site provided an unbalanced sampling structure with four times as many reference site samples.

% Difference	Power	Total N	Exposed Site N
10	0.80	670	134
	0.85	770	154
	0.90	900	180
	0.95	1110	222
15	0.80	310	62
	0.85	350	70
	0.90	410	82
	0.95	505	101
20	0.80	175	35
	0.85	200	40
	0.90	235	47
	0.95	290	58
25	0.80	115	23
	0.85	130	26
	0.90	155	31
	0.95	190	38

Table 5. Sample sizes required for detecting differences in log p,p'DDE concentrations between a site of suspected exposure and a reference site provided a balanced sampling structure.

% Difference	Power	Total N	Exposed Site N
10	0.80	272	136
	0.85	310	155
	0.90	364	182
	0.95	448	224
15	0.80	122	61
	0.85	140	70
	0.90	164	82
	0.95	202	101
20	0.80	70	35
	0.85	80	40
	0.90	94	47
	0.95	114	57
25	0.80	46	23
	0.85	52	26
	0.90	60	30
	0.95	74	37

Table 6. Sample sizes required for detecting differences in log p,p'DDE concentrations between a site of suspected exposure and a reference site provided an unbalanced sampling structure with twice as many reference site samples.

% Difference	Power	Total N	Exposed Site N
10	0.80	306	102
	0.85	351	117
	0.90	408	136
	0.95	504	168
15	0.80	138	46
	0.85	156	52
	0.90	183	61
	0.95	225	75
20	0.80	78	26
	0.85	90	30
	0.90	105	35
	0.95	129	43
25	0.80	51	17
	0.85	60	20
	0.90	69	23
	0.95	84	28

Table 7. Sample sizes required for detecting differences in log p,p'DDE concentrations between a site of suspected exposure and a reference site provided an unbalanced sampling structure with three times as many reference site samples.

% Difference	Power	Total N	Exposed Site N
10	0.80	364	91
	0.85	416	104
	0.90	484	121
	0.95	600	150
15	0.80	164	41
	0.85	188	47
	0.90	216	54
	0.95	268	67
20	0.80	92	23
	0.85	108	27
	0.90	124	31
	0.95	152	38
25	0.80	60	15
	0.85	68	17
	0.90	80	20
	0.95	100	25

Table 8. Sample sizes required for detecting differences in log p,p'DDE concentrations between a site of suspected exposure and a reference site provided an unbalanced sampling structure with four times as many reference site samples.

% Difference	Power	Total N	Exposed Site N
10	0.80	425	85
	0.85	485	97
	0.90	570	114
	0.95	700	140
15	0.80	190	38
	0.85	220	44
	0.90	255	51
	0.95	315	63
20	0.80	110	22
	0.85	125	25
	0.90	145	29
	0.95	180	36
25	0.80	70	14
	0.85	80	16
	0.90	95	19
	0.95	115	23

Discussion

The purpose of this investigation was to determine the sample sizes that would be necessary to detect regionally elevated contaminant concentrations when compared to reference site concentrations. This highlights one of the benefits of maintaining an ongoing contaminant monitoring program. Having extensive records of background contaminant concentrations can provide an invaluable reference for identifying new areas of concern. The contaminants discussed here are persistent and move through the aquatic food web. This makes individual watersheds (HUCs) a logical fine scale focus of trend investigation. In addition, humans share a place at the top of that food web with bald eagles, which makes monitoring bald eagle exposure a useful indicator of safety for humans. It is important to keep power in mind when collecting and analyzing data. Especially in the environmental sciences, questions of compromise between Type I and Type II error must be weighed against the cost and consequences of each (Buhl-Mortensen, 1996; Fairweather, 1991).

The sample sizes necessary to meet the power objectives for 10% or 15% changes in contaminant concentration would be difficult to achieve within a single year for a single watershed. Many HUCs in the state have only a limited number of samples taken each year. Roe (2004), however, showed that neighboring HUCs could be combined in order to achieve a sufficiently large sample size. Furthermore, if no samples are available for adjacent HUCs within a sampling year, they may be grouped with adjacent HUCs from different sampling years. In the case of the kind of monitoring described here,

caution should be taken not to use data from outside the area of suspected exposure. While grouping adjacent HUCs from different sampling years may allow for greater sample sizes, it may also obfuscate locally elevated contaminant concentrations that were the intended focus of detection. The sample sizes necessary to meet power objectives for the detection of 20% or 25% elevation in contaminant concentration are not prohibitively large, and could likely be met within the normal sampling structure of the Michigan BEBP.

Sample sizes necessary for detection at every level are smaller for p,p'DDE concentrations than for PCB concentrations. This is a natural consequence of the observed lower variance in measured concentrations for p,p'DDE. Our estimates of variance represent the 'noise' in the data that must be overcome to detect a statistically significant signal. The log-scale standard deviation for PCB concentrations was 1.23 ppb, while the log-scale standard deviation estimate for p,p'DDE was 1.07, which suggests more variability exists in the distribution of PCBs and therefore, a greater sampling effort is required to provide standard errors sufficiently small to result in statistical significance.

A key part of Michigan's environmental quality monitoring program was the timing of sample collection, analyses, and reporting of the monitoring data for each HUC watershed. The strategy was intended to provided data concerning contaminant levels for the HUCs prior to the initiation of the National Pollutant Discharge Elimination System permit development and renewal process (MDEQ, 1997). The BEBP was therefore developed on a five-year watershed cycle that allows for the HUCs to be monitored two

to three years prior to the actual permit issuance year. For this reason, it is also important to exhibit caution when combining neighboring HUCs, to ensure that all data collected for a HUC will be analyzed and available before that HUCs sampling year. This must be done to make sure that in cases of suspected elevated exposure for a locale, the permit process is not conducted before data are reported.

In conclusion, this analysis has shown that data from the Michigan BEBP could provide a valuable resource for documenting areas of concern in the state. With sufficient sample sizes to detect 20% and 25% increases in contaminant concentration with a power of 0.80 or 0.85 easily obtainable and a large available pool of reference site samples for comparison, these data could help identify watersheds with emerging contaminant problems. If the area of suspected elevated contaminant concentrations was large enough that the combining of neighboring watersheds is appropriate, then greater power or smaller shifts in contaminant concentration could be detected.

Literature Cited

Bowerman, W. W. (1993). *Regulation of bald eagles (Haliaeetus leucocephalus) productivity in the Great Lakes Basin: And ecological and toxicological approach.* (PhD), Michigan State University, East Lansing, MI, USA.

Bowerman, W. W., Best, D. A., Giesy, J. P., Shieldcastle, M. C., Meyer, M. W., Postupalsky, S., & Sikarskie, J. G. (2003). Associations between regional differences in polychlorinated biphenyls and dichlorodiphenyldichloroethylene in blood of nestling bald eagles and reproductive productivity. *Environmental Toxicology and Chemistry, 22*(2), 371-376.

Bowerman, W. W., Best, D. A., Grubb, T. G., Zimmerman, G. M., & Giesy, J. P. (1998). Trends of contaminants and effects in bald eagles of the Great Lakes basin. *Environmental Monitoring and Assessment, 53*(1), 197-212.

Bowerman, W. W., Giesy, J. P., Best, D. A., & Kramer, V. J. (1995). A REVIEW OF FACTORS AFFECTING PRODUCTIVITY OF BALD EAGLES IN THE GREAT-LAKES REGION - IMPLICATIONS FOR RECOVERY. *Environmental Health Perspectives, 103*, 51-59.

Bowerman, W. W. I. V., Best, D. A., Giesy, J. P., Jr., Kubiak, T. J., Sikarskie, J. G., Meyburg, B. U., & Chancellor, R. D. (1994). The influence of environmental contaminants on bald eagle Haliaeetus leucocephalus populations in the Laurentian Great Lakes, North America. *Raptor conservation today.*, 703-707.

Buehler, D. A. (2000). Bald Eagle: Haliaeetus leucocephalus. *Birds of North America*(506), 1-40.

Buhl-Mortensen, L. (1996). Type-II statistical errors in environmental science and the precautionary principle. *Marine Pollution Bulletin, 32*(7), 528-531.

Cope W. G., L., R. B. , and Hodgson, E. . (2004). *Classes of Toxicants: Use Classes. A textbook of modern toxicology.* New Jersy: John Wiley & Sons: 49-73.

Dykstra, C. R., Meyer, M. W., Stromborg, K. L., Warnke, D. K., Bowerman, W. W., & Best, D. A. (2001). Association of low reproductive rates and high contaminant levels in bald eagles on Green Bay, Lake Michigan. *Journal of Great Lakes Research, 27*(2), 239-251.

Fairweather, P. G. (1991). STATISTICAL POWER AND DESIGN REQUIREMENTS FOR ENVIRONMENTAL MONITORING. *Australian Journal of Marine and Freshwater Research, 42*(5), 555-567.

Farrell, P. J., & Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation, 76*(9), 803-816. doi: 10.1080/10629360500109023

Giesy, J. P., Bowerman, W. W., Mora, M. A., Verbrugge, D. A., Othoudt, R. A., Newsted, J. L., . . . Tillitt, D. E. (1995). Contaminants in fishes from great lakes-influenced sections and above dams of three Michigan Rivers: III. Implications for health of bald eagles. *Archives of Environmental Contamination and Toxicology, 29*(3), 309-321.

Grubb, T. G., Wiemeyer, S. N., & Kiff, L. F. (1990). EGGSHELL THINNING AND CONTAMINANT LEVELS IN BALD EAGLE EGGS FROM ARIZONA, 1977 TO 1985. *Southwestern Naturalist, 35*(3), 298-301.

Hoffman, D. J., Melancon, M. J., Klein, P. N., Rice, C. P., Eisemann, J. D., Hines, R. K., . . . Pendleton, G. W. (1996). Developmental toxicity of PCB 126 (3,3',4,4',5-pentachlorobiphenyl) in nestling American kestrels (Falco sparverius). *Fundamental and Applied Toxicology, 34*(2), 188-200.

Leith, K. F., Bowerman, W. W., Wierda, M. R., Best, D. A., Grubb, T. G., & Sikarske, J. G. (2010). A comparison of techniques for assessing central tendency in left-censored data using PCB and p,p'DDE contaminant concentrations from Michigan's Bald Eagle Biosentinel Program. *Chemosphere, 80*(1), 7-12.

MDEQ. (1997). A strategic environmental quality monitoring program for Michigan's surface waters. *MI/DEQ/SWQ-96/152* (MI/DEQ/SWQ-96/152 ed., Vol. January 1997). Lansing, Michigan: Michigan Department of Environmental Quality.

Noughabi, H. A., & Arghami, N. R. (2010). Monte Carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation, 81*(8), 965-972. doi: 10.1080/00949650903580047

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33.

Roe, A. (2004). *SPATIAL AND TEMPORAL ANALYSES OF ENVIRONMENTAL CONTAMINANTS AND TROPHIC STATUS OF BALD EAGLES IN THE GREAT LAKES REGION.* (PhD), Clemson University, Clemson.

Romão, X., Delgado, R., & Costa, A. (2009). An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation, 80*(5), 545-591. doi: 10.1080/00949650902740824

SAS Institute Inc. (2000-2004). SAS 9.1.2 Help and Documentation. Cary, NC: SAS Institute Inc.

Seier, E. (2002). Comparison of tests for univariate normality. *Interstat, 1*, 1-17.

Wiemeyer, S. N., Bunck, C. M., & Stafford, C. J. (1993). ENVIRONMENTAL CONTAMINANTS IN BALD EAGLE EGGS 1980-84 AND FURTHER INTERPRETATIONS OF RELATIONSHIPS TO PRODUCTIVITY AND SHELL THICKNESS. *Archives of Environmental Contamination and Toxicology, 24*(2), 213-227.

Wiemeyer, S. N., Lamont, T. G., Bunck, C. M., Sindelar, C. R., Gramlich, F. J., Fraser, J. D., & Byrd, M. A. (1984). ORGANOCHLORINE PESTICIDE, POLYCHLOROBIPHENYL, AND MERCURY RESIDUES IN BALD EAGLE EGGS - 1969-79 - AND THEIR RELATIONSHIPS TO SHELL THINNING AND REPRODUCTION. *Archives of Environmental Contamination and Toxicology, 13*(5), 529-549.

Wiemeyer, S. N., Lamont, T. G., Bunck, C. M., Sindelar, C. R., Gramlich, F. J., Fraser, J. D., & Byrd, M. A. (1984). Organochlorine pesticide, polychlorobiphenyl, and mercury residues in bald eagle eggs—1969–79—and their relationships to shell thinning and reproduction. *Archives of Environmental Contamination and Toxicology, 13*(5), 529-549.

Wierda, M. R. (2009). *Using bald eagles to track spatial and temporal trends of contaminants in Michigan's aquatic systems.* (PhD Wildlife and Fisheries Biology Dissertation), Clemson University, Clemson, SC.

Chapter 2: A comparison of techniques for assessing central tendency in left-censored data using PCB and p,p'DDE contaminant concentrations from Michigan's Bald Eagle Biosentinel Program.

Introduction

Monitoring contaminant levels in the environment is an important part of understanding the fate of ecosystems after a chemical insult. As levels decrease over time, limitations in analytical equipment create a lower bound below which contaminant levels cannot be accurately reported. This results in datasets with observations below the detection limit (DL) that are reported only as 'non-detect' or '< DL' and no value is provided. This type of distribution is called 'left-censored', as the low-end observations that are unknown generally occur near the origin of the x-axis in figures. Several options for the analysis of these datasets have been investigated leading to the conclusion that methods replacing all non-detects with a single value (substitution methods) are frequently inferior (Antweiler & Taylor, 2008; Baccarelli et al., 2005; Eastoe et al., 2006; Helsel, 2006, 2005b, 2005a; Liu et al., 1997; Needham et al., 2007; Singh & Nocerino, 2002). Specifically, it has been shown that the bias caused by substitution increases dramatically as the percent of observations censored increases (Eastoe et al., 2006). In spite of this, various substitution methods continue to be used in research, frequently with little regard for the proportion of observations censored.

In addition to censored observations, another complication when analyzing environmental contaminant data is the possibility that datasets display a right skew. This

occurs when a few samples show very high concentrations while the general tendency is for concentrations to be lower. This distribution is common in environmental data and can frequently be accommodated by log-transformation. The median has traditionally been an accepted measure of central tendency for data which do not fit a normal distribution well, and this approach has been used in almost every field of scientific inquiry. Focus has shifted to newer approaches as more complex methods have been developed and computing power has grown to make them feasible for the average researcher. While the median is still useful in that it is not based on indefensible assumptions about the shape of the distribution, it does not make use of all the information contained in a dataset. A final complexity is added by the fact that lognormal data are frequently summarized using the geometric mean, which is particularly sensitive to the choice of substitution value.

In 1997, the Michigan Department of Environmental Quality (MDEQ) implemented a Bald Eagle Biosentinel Program (BEBP) to monitor trends of a suite of organic pollutants under the Clean Michigan Initiative (MDEQ, 1997). The data analyzed here are 234 observations of polychlorinated biphenol (PCB) and p,p'-Dichlorodiphenyldichloro-ethylene (p,p'DDE) concentrations found in nestling bald eagle plasma samples from throughout the State of Michigan. Current statistical methods include tests of significant differences between regions at several geographic scales, and calculating descriptive statistics in the form of geometric means. Substitution is currently used in cases of non-detect, or left-censored, observations.

This choice of methods for addressing the non-detects does not affect testing of significant differences between regions. The monitoring program uses nonparametric Kruskal-Wallis and Wilcoxon tests which are rank-based and make no assumptions about distribution and so, are not sensitive to the problems of substitution (Helsel, 2005). However, summary statistics are reported as geometric means, which are affected by the choice of substituted value. The BEBP program currently substitutes the near-zero value of '0.0001' for concentrations at non-detectable levels. This near-zero value might appear to have little influence on the resulting calculations to those accustomed to arithmetic mean calculation because the arithmetic mean is a function of addition, for which '0' is the identity value. Geometric means on the other hand are a function of multiplication, for which the identity value is '1', while near-zero values (like '0.0001) have a drastic impact on the product.

The desire to summarize data more accurately has fueled recent comparisons of proposed analytical alternatives to substitution or simple median reporting (Helsel, 2005). One alternative, maximum-likelihood estimation (MLE), forces the researcher to assume the shape of the underlying distribution, but is powerful if this assumption is correct. These MLE methods have been explored in a variety of environmental applications (Helsel, 2006, 2005b; Antweiler & Taylor, 2008; Jain et al., 2008; Singh & Nocerino 2002). Kaplan-Meier (KM) estimators have also been proposed. This estimator began as a nonparametric method of estimating the central tendency in right-censored survival data, but is gaining popularity for left-censored datasets. Among those who have explored the application of KM calculations in left-censored environmental data are

Antweiler and Taylor (2008), Eastoe et al. (2006), and Helsel, (Helsel, 2005b; 2005a). Multiple imputation (MI) has been proposed as a 'fill-in' technique that can be used to first estimate an appropriate distribution shape based on uncensored values, then samples from the values that would be found in the censored tail. This technique has been addressed in comparison studies by Antweiler & Taylor (2008), Baccarelli et al. (2005), Eastoe et al. (2006), Helsel (Helsel, 2005b; 2005a), Krishnamoorthy et al. (2009), and Singh & Nocerino (2002). Right-skewed, left-censored data were the focus of Singh & Nocerino (2002), who applied many analysis techniques common for left-censored data and assessed their performance when the observations also displayed a right skew. They found that left-censored datasets were more difficult to accurately summarize in the presence of a right skew.

Few case studies have been published and substitution is still in wide use (Baccarelli et al., 2005; Eastoe et al., 2006). This study explored the effects of non-detect data and their treatment on summary statistics. The data analyzed in this paper represent both large (N=234) and moderate (n=12 to n=64) sample sizes with both good and marginal fit with a log transformation. Summary statistics were calculated using the current method of substitution with '0.0001', the common method of substitution with '½*DL', MI, and KM estimation. The median was also calculated for comparison with all four methods. The objectives were to (1) assess the fit of lognormal distribution to the data, (2) compare and contrast the performance of the four methods of non-detect handling in terms of estimated geometric mean, comparison to the median, and standard error, and (3) make recommendations based on those results.

Methods

Concentrations of total PCBs and p,p'DDE ($^{\mu g}/_{kg}$ ww) in plasma collected from nestling bald eagles across Michigan from 1999 to 2003 were used in these analyses. In addition to analysis as a single dataset representing the whole State of Michigan, data were also classified geographically by subpopulation, based on the classifications used in the BEBP. Subpopulations were defined by first subdividing the state spatially into the categories of Great Lakes and Inland breeding areas. Great Lakes breeding areas are defined as being within 8.0 km of Great Lakes shorelines and/or along tributaries open to Great Lakes fish runs and inland breeding areas are defined as being greater than 8.0 km from the Great Lakes shorelines and not along tributaries open to Great Lakes fish runs. These categories are then further subdivided into four Great Lakes and two Inland groups. The Great Lakes subpopulations consisted of Lake Superior (LS), Lake Michigan (LM), Lake Huron (LH), and Lake Erie (LE). The Inland subpopulations consisted of Upper Peninsula (UP), and Lower Peninsula (LP) (Wierda, 2009). The data analysis for this paper was generated using SAS® software, Version 9.1.2 of the SAS system for Windows. Copyright 2000-2004 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Assessment of Fit

Data were analyzed for significant departure from the lognormal distribution using the UNIVARIATE PROCEDURE with the options 'normal' and 'plot' activated (SAS Institute Inc., 2000-2004). This procedure produces several fit statistics. The

Kolomogorov-Smirnov statistic was used to assess the assumption of fit to the lognormal distribution. The fit was classified as 'good' with p-values ≥ 0.05, 'marginal' with p-values between 0.01 and 0.05, and fit was rejected for p-values<0.01.

Geometric Means and Standard Error Calculation

Geometric means and standard errors were calculated for all of the proposed methods. The median was used for comparison and obtained from the univariate analysis discussed above. For substitution methods, geometric means and standard errors were calculated by log-converting observations, calculating the mean and standard error of the transformed data using the MEANS PROCEDURE (SAS Institute Inc., 2000-2004), and then converting back to the original scale. Monte Carlo simulations were run in order to test the significance of the divergence of the geometric mean (using each method of substitution) from the median. Each of these simulations resulted in a 'p-value' representing the probability of the observed divergence occurring due to sampling error alone. Simulation resulting in p-values of 0.05 or less were considered evidence of a significant substitution method effect.

Geometric means and errors were calculated using the multiple imputation methods based on those described in Krishnamoorthy et al. (2009) and the MI PROCEDURE(SAS Institute Inc., 2000-2004). Ten imputations were used on the recommendation of Jain, et al. (2008). The option 'EM' was used to implement the maximum likelihood method of adjusting the approximated distribution from which imputed values were drawn. Bounds were set to ensure that no negative values were imputed.

Kaplan-Meier estimates of geometric mean and standard error were calculated using the LIFETEST PROCEDURE (SAS Institute Inc., 2000-2004) on log-transformed data. This procedure is designed to perform survival analysis for right censored data, so data were transformed to reflect a right censored distribution. The transformation was conducted by subtracting all log-transformed observations from a number larger than the largest observation. This was done for PCBs by subtracting all observations from 12, and for p,p'DDE by subtracting all observations from 10. Results were then transformed back to reflect geometric means and standard errors in the original units.

Results

Fit of Data to Lognormal Distribution

Kolmogorov-Smirnov (KS) tests did not force us to reject the assumption of lognormal distribution in all cases, but did suggest significant non-normality in others. Tests conducted for PCBs and p,p'DDEs at both the whole state and subpopulation level resulted in different conclusions.

At the state level, the PCB distribution was classified as marginal ($P=0.0206$) and p,p'DDE distribution was classified as significantly differing from lognormal ($P<0.01$). This was likely due to the presence of six moderate outliers in the upper tail of the p,p'DDE distribution. When the outliers were removed, the KS test resulted in no evidence for significant departure from lognormality ($P=0.1338$), suggesting a good fit.

When broken down into the geographical units of subpopulation, KS analysis of PCB concentrations suggested that Lake Erie and Michigan coastal regions exhibited

marginal evidence for departure from the lognormal distribution. Lake Huron and Superior coastal regions and both Upper and Lower Peninsula inland regions showed no significant departure from the lognormal distribution for PCB concentrations, suggesting a good fit of the data. For p,p'DDE concentrations, the Lake Erie coastal region showed significant departure from normality ($P<0.01$), but all other regions showed good fit and the assumption of a lognormal distribution was considered sound.

Geometric Means, Medians, and Standard Errors

For comparisons made at the whole state level, measures of central tendency in PCBs ranged from 33 $^{\mu g}/_{kg}$ for the current method, to 78 $^{\mu g}/_{kg}$ using MI. The median PCB concentration was 77 $^{\mu g}/_{kg}$. For p,p'DDE, central tendency measures ranged from 6 $^{\mu g}/_{kg}$ using the current method to 20 $^{\mu g}/_{kg}$ using MI, with a median concentration of 17 $^{\mu g}/_{kg}$. In both cases the MI method produced the highest estimate of geometric mean, but was near the median and KM estimate, which was 69 $^{\mu g}/_{kg}$ for PCBs and 18 $^{\mu g}/_{kg}$ for p,p'DDE. Comparisons for both contaminants also resulted in the lowest estimate of geometric mean using the current method, as would be expected based on the mathematical underpinnings of geometric mean calculation. The method of substitution using half the detection limit was consistently lower than the MI, K-M, and median, but much closer than the current substitution method. Geometric means for each of the methods discussed as well as the median are shown for PCBs and p,p'DDE in Figure 1. In addition to summary statistics, the figure shows error bars representing one standard error above and below the geometric mean for each method and each contaminant.

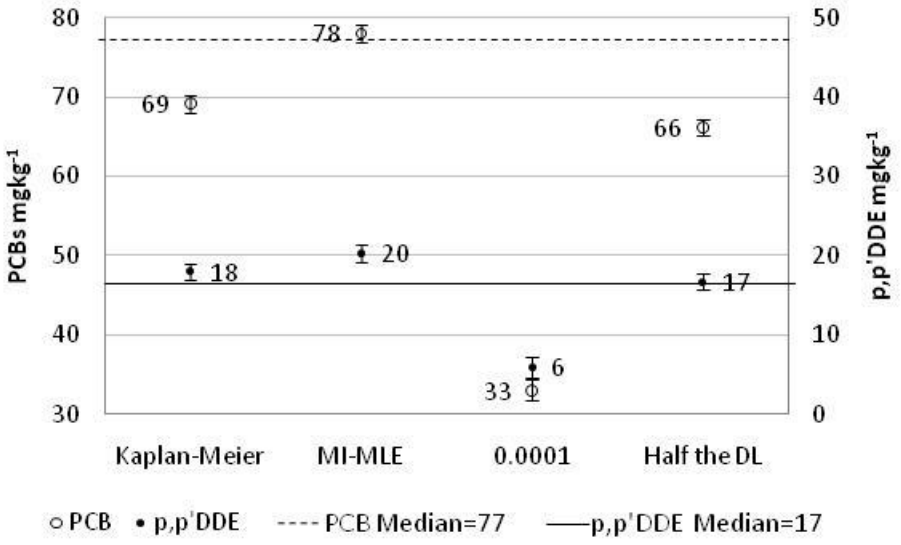

Figure 1. Geometric means +/- 1 SE resulting from four methods of calculation for PCB and p,p'DDE levels within the state of Michigan. Median is included for comparison. These 234 observations represent samples collected from 1999-2003.

The standard errors of the geometric mean at the state level for all of the methods discussed were similar. Table 1 displays the standard errors, number of observations (N) and the rate of censorship for each contaminant. For PCB concentrations, measures of standard error ranged from 1.1 using Kaplan-Meier, MI and half the DL methods to 1.3 using the current 0.0001 substitution value. For p,p'DDE concentrations, measures of standard error also ranged from 1.1 to 1.3. Again, the current method of substitution was higher, with the remaining three methods lower and in agreement.

Table 1. Shows the standard error of the geometric mean resulting from four methods of calculation and the rate of censorship for the entire state of Michigan for PCBs and p,p'DDE. The 234 observations represent samples collected from 1999-2003.

Error Method	PCB (μgkg^{-1})	DDE (μgkg^{-1})
	6.41% Censored	10.26% Censored
Kaplan-Meier	1.1	1.1
MI-MLE	1.1	1.1
0.0001	1.3	1.3
Half the DL	1.1	1.1

Results at the subpopulation level follow the same trend as the whole state, with the current substitution method producing depressed geometric mean estimates and elevated standard errors. The results of the geometric mean analysis and the medians for PCB and p,p'DDE concentrations are summarized in Table 2. The standard errors for PCB and p,p'DDE concentrations are provided in Table 3. Tables 2 and 3 also display the number of observations (n) and rate of censorship for each subpopulation. Subpopulations that were omitted in these tables were those that had no censored values. In all cases with no censored values the KM method resulted in the same estimate as both substitution methods. Because these were instances in which no observations were missing and so, no substitutions were made, they were omitted to prevent them from being inappropriately interpreted as an instance of agreement between KM and substitution methods.

Table 2. Medians and geometric means of total PCBs and p'p,DDE concentrations for each subpopulation and each method of calculation. Also included are the number of observations and rate of censorship for each subpopulation.

	PCB ($\mu g kg^{-1}$)							**p,p'DDE** ($\mu g kg^{-1}$)						
Subpopulation	Median	Kaplan-Meier	MI-MLE	Current 0.0001	Half the DL	n	Censorship (%)	Median	Kaplan-Meier	MI-MLE	Current 0.0001	Half the DL	n	Censorship (%)
Lake Huron	135	121	124	$43^{0.0224}$	105	12	8	36	29	29	$11^{0.0001}$	$25^{0.0182}$	12	8
Lake Superior	-	-	-	-	-	-	-	32	29	33	8^{*}	25	45	11
Lower Peninsula	33	33	35	16^{*}	31	49	6	12	11	12	2^{*}	$10^{0.0218}$	49	14
Upper Peninsula	36	$29^{0.0077}$	36	4^{*}	$26^{0.0004}$	64	17	13	12	14	2^{*}	$11^{0.0306}$	64	17

Superscript P-values are provided for estimates that showed a significant substitution method effect based on Monte Carlo simulations.
*P-value was less than 0.0001.
There were no left-censored observations for PCB in the Lake Superior subpopulation.

Table 3. Standard errors of the geometric means of total PCBs and p'p,DDE concentrations for each subpopulation and each method of calculation. Also included are the number of observations and rate of censorship for each subpopulation.

	PCB ($\mu g kg^{-1}$)						**p,p'DDE** ($\mu g kg^{-1}$)					
Subpopulation	Kaplan-Meier	MI-MLE	Current 0.0001	Half the DL	n	Censorship (%)	Kaplan-Meier	MI-MLE	Current 0.0001	Half the DL	n	Censorship (%)
Lake Huron	1.4	1.4	3.4	1.5	12	8	1.2	1.2	2.9	1.3	12	8
Lake Superior	*	*	*	*	*	*	1.2	1.1	1.8	1.2	45	11
Lower Peninsula	1.1	1.1	1.6	1.1	49	6	1.1	1.1	1.8	1.1	49	14
Upper Peninsula	1.1	1.1	1.9	1.1	64	17	1.1	1.1	1.8	1.1	64	17

*There were no left-censored observations for PCB in the Lake Superior subpopulation.

Discussion

Fit of Data to Lognormal Distribution

While not all of the datasets analyzed conformed well to the lognormal distribution, most were either a marginal or good fit. The lognormal distribution is common when handling environmental data, so it is not surprising that many of these contaminant distributions are well approximated by it. Singh and Nocerino (2002) have cautioned that parametric methods of calculating central tendency measures in left-censored datasets can be less reliable when observations do not fit the assumed distribution. In this analysis, the MI method would have been vulnerable to such problems. In all cases, however, the MI geometric mean estimate was very close to the KM estimate, which is not vulnerable to such parametric assumptions. Likewise, in the calculation of standard errors, MI performed almost identically to the KM method. This suggests that MI is robust to at least minor deviations from an assumed distribution.

Geometric Means, Medians, & Standard Errors

This study shows that the current method of near-zero substitution for calculating geometric means with left-censored data and a right skew performs poorly relative to the methods used for comparison in this study. It resulted in the lowest estimated central tendency, which was farthest from the median, and resulted in the highest estimated standard error when compared to other methods. Several investigations have provided evidence that methods replacing all non-detects with a single value (substitution methods) can introduce bias (Antweiler & Taylor, 2008; Baccarelli et al., 2005; Eastoe et al., 2006; D. R. Helsel, 2005; Dennis R. Helsel, 2005; Helsel, 2006; Liu et al., 1997;

Needham et al., 2007; Singh & Nocerino, 2002). This is especially true for datasets with a large proportion of censored observations because bias caused by substitution increases dramatically as the percent of observations censored increases (Eastoe et al., 2006). While several of the alternatives are analytically intensive, many statistical packages now recognize the need and are designed to conduct such analyses. As more programs accommodate this need the programming skill required will no longer be a prohibitive factor. The problems of substitution in general, are compounded by the use of geometric means and especially the choice of substituted value (Helsel, 2006; 2005b). While it may seem appropriate to choose a near-zero value such as '0.0001', this inference is based on the mathematical underpinnings of the arithmetic mean, which differ from those of geometric mean calculation. Arithmetic mean calculation is governed by the properties of addition, for which '0' is the identity value. This means that '0' is the number which can be added to a series without changing the sum. In calculating the arithmetic mean, a '0' allows the sum to remain unchanged while increasing *N*, which is the divisor. In this regard, the purpose of substitution is to serve as a place holder that lets *N* increase without changing the numerator, thereby allowing the non-detects to affect the quotient only by inflating *N*. Geometric mean calculation, however, is governed by the properties of multiplication, for which '1' is the identity value. In multiplication, in contrast to addition, near-zero values have a dramatic effect, while values near one are of the lowest impact.

Imagine the difference in this simple example, between the effects of:

(1) 10,000 + 0.0001 = 10,000.0001

(2) 10,000 * 0.0001 = 1

When adding in (1), the result is very near the number with which we began. When multiplying in (2) though, the result is 1, which is a drastic decrease from 10,000.

Estimates of geometric mean and standard errors for all methods except the current substitution method were largely in agreement. This includes the other substitution method tested here of '½*DL', which is common practice for contaminant monitoring programs. In addition, these estimates were in close agreement with the median, which suggests that they are capturing the central tendency of contaminant concentrations and not overly sensitive to the censorship or skew in the dataset. The maximum likelihood based MI estimates of geometric mean were consistently highest when accommodating these datasets, which suggests that they are the most vulnerable to right skew of the methods considered here. Indeed, in Singh and Nocerino's (2002) discussion of handling censored data in the presence of a right skew, they warned that such distribution based methods were "particularly susceptible to problems caused by outliers." They concluded that for large sample sizes and only when distributions could be satisfactorily fit, MLE-based analyses were good alternatives. As stated above, some of our data were shown to be a poor fit to the assumed distribution. However, MI provides an advantage over strict maximum likelihood estimators in that when it is used to 'fill-in' missing observations, sampling is done *multiple* times. This provides the

distinct advantage of estimating the variance resulting from the procedure itself versus the variability in the actual contaminant concentrations. Mutiple imputation estimators have also been previously found to produce unbiased estimates when the proportion of uncensored values was less than 50% (Jain et al., 2008). Multiple-imputation estimates of standard error were similar to estimates produced by all but the current substitution method.

Kaplan-Meier estimates were first derived as a way of determining mean survival in datasets in which not all members of the sample died at the end of the experiment. This resulted in right censored distributions, which are common in engineering and medical trials (Helsel, 2005). Increasingly, the common problems in analyzing right and left censored data have drawn researchers in the environmental field to apply these techniques. Originally, left-censored data were simply transformed to make a right-censored distribution by subtracting them from an arbitrary value larger than the largest observation. As this method grows in acceptance, programs have begun to accommodate left-censored data without such transformations.

In this study, KM estimates did not seem as sensitive to the effects of right-skewed data, which is a major benefit of using a nonparametric analysis technique. The KM estimates of both geometric mean and standard error were overall quite similar to those produced by all but the current substitution method, though the geometric means estimates were consistently lower than the MI estimates, where differences occurred. In other comparisons of data handling methods for left censored datasets, KM was determined to perform best in the determination of summary statistics (Antweiler &

Taylor, 2008). As a testament to its robustness, it has been used as the standard of comparison in other studies of left censored data (Eastoe et al., 2006).

Recommendations

Based on the findings here, KM statistics provide the best estimates of geometric means in data with both left hand censorship and a right skew, like those generated by the Michigan BEBP. Differences between KM estimates and MI estimates were minor, which may tempt the conclusion that they are equally valid for these data. However, based on the subtle trend here of MI to be pulled upward, and published evidence of the tendency of parametric analyses like MI to be biased by skewed data (Eastoe et al., 2006), KM seems the best option. Use of the nonparametric KM also provides theoretical consistency in that significant difference testing is already performed using nonparametric techniques.

For the dataset in this study, the common practice of substitution with '½*DL' resulted in estimates of both geometric means and standard errors that did not differ greatly from other methods compared. This must be interpreted with caution, however, since these data had low rates of censorship (6.41% for PCBs and 10.26% for p,p'DDE, at the state level). It has been shown that the bias caused by substitution increases dramatically as the percent of observations censored increases (Eastoe et al., 2006). The agreement between this substitution method and more complex methods is likely a reflection of the low levels of substitution in these data; it should not be interpreted as evidence of equivalence between the '½*DL' substitution method and MI or KM methods. It may be concluded that substitution of '½*DL' would be an acceptable

treatment of censored values *only* for studies with low levels of censorship. Substitution is still common practice in toxicological studies, which makes it tempting to employ for the purpose of consistency. However, substitution will become an increasingly problematic solution as monitored contaminants become less prevalent and a larger proportion of samples contain contaminant levels in the nondetectable range. This is evident when comparing the data presented here to historical data. For example, in the years from 1987-1993 Lake Huron nestlings provided no samples with non-detectable PCB concentrations compared with 8% of samples with non-detectable PCB concentrations in these data. Only 4% of samples from Lower Peninsula nestlings and only 9% of samples from the Upper Peninsula had non-detectable PCB concentrations in the 1987-1993 dataset, compared with 6% and 17%, respectively here (Bowerman, 1993). We believe that as more studies of this nature are published and software increasingly accommodates left censored data, substitution methods will become less prevalent.

Literature Cited

Antweiler, R. C., & Taylor, H. E. (2008). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. *Environmental Science & Technology*, *42*(10), 3732-3738.

Baccarelli, A., Pfeiffer, R., Consonni, D., Pesatori, A. C., Bonzini, M., Patterson, D. G., Bertazzi, P. A., Landi, M. T. (2005). Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the Seveso chloracne study. *Chemosphere*, *60*(7), 898-906.

Bowerman, W. W. (1993). *Regulation of bald eagles (Haliaeetus leucocephalus) productivity in the Great Lakes Basin: And ecological and toxicological approach.* Unpublished PhD dissertation, Michigan State University, East Lansing, MI, USA.

Eastoe, E. F., Halsall, C. J., Heffernan, J. E., & Hung, H. (2006). A statistical comparison of survival and replacement analyses for the use of censored data in a contaminant air database: A case study from the Canadian Arctic. *Atmospheric Environment*, *40*(34), 6528-6540.

Helsel, D. R. (2005). More than obvious: Better methods for interpreting nondetect data. *Environmental Science & Technology*, *39*(20), 419A-423A.

Helsel, D. R. (2005). *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. Hoboken: John Wiley & Sons, Inc.

Helsel, D. R. (2006). Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*, *65*(11), 2434-2439.

Jain, R. B., Caudill, S. P., Wang, R. Y., & Monsell, E. (2008). Evaluation of maximum likelihood procedures to estimate left censored observations. *Analytical Chemistry, 80*(4), 1124-1132.

Krishnamoorthy, K., Mallick, A., & Mathew, T. (2009). Model-Based Imputation Approach for Data Analysis in the Presence of Non-detects. *Annals of Occupational Hygiene, 53*(3), 249-263.

Liu, S., Lu, J.-C., Kolpin, D. W., & Meeker, W. Q. (1997). Analysis of Environmental Data with Censored Observations. *Environmental Science & Technology, 31*(12), 3358-3362.

MDEQ. (1997). A strategic environmental quality monitoring program for Michigan's surface waters. Vol. January 1997. Michigan Department of Environmental Quality, Lansing, Michigan.

Needham, L. L., Naiman, D. Q., Patterson, D. G., & LaKind, J. S. (2007). Assigning concentration values for dioxin and furan congeners in human serum when measurements are below limits of detection: An observational approach. *Chemosphere, 67*(3), 439-447.

SAS Institute Inc. (2000-2004). SAS 9.1.2 Help and Documentation. Cary, NC: SAS Institute Inc.

Singh, A., & Nocerino, J. (2002). Robust estimation of mean and variance using environmental data sets with below detection limit observations. *Chemometrics and Intelligent Laboratory Systems, 60*(1-2), 69-86.

Wierda, M. R. (2009). *Using bald eagles to track spatial and temporal trends of contaminants in Michigan's aquatic systems.* Unpublished PhD dissertation, Clemson University, Clemson, SC.

Chapter 3: Assessment of the effects of model specification on inferential conclusions regarding reproductive outcome as a function of site classification.

Introduction

Background

The bald eagle *(Haliaeetus leucocephalus)* is considered to be primarily piscivorous, but also opportunistically forages on an array of avian, mammalian, and reptilian prey. They are associated with aquatic habitats throughout North America including coastal areas, rivers, lakes, reservoirs, and forested shorelines (Buehler, 2000). Mean clutch size has been estimated at 1.87 eggs per clutch and clutches usually range from one to three eggs (Stalmaster, 1987). Extensive research has been conducted on this high-profile raptor addressing life history characteristics and the influences of various stressors on reproduction. For this reason, tracking reproductive outcomes is useful both as an indicator of the health of the population itself, and as broad scale indicator of ecosystem changes.

It has also been shown through careful observation by ornithologists, that there is regional variation of recovery within eagle populations. It appeared that nesting pairs along the Great Lakes coast were rebounding less successfully than inland populations (Best et al., 1994). This suspicion was later confirmed as studies of contaminant concentrations from collected eggs, and later from nestlings, showed higher contaminant concentrations in areas with a Great Lakes centered prey base when compared to inland

areas (Wiemeyer et al., 1984; Bowerman et al., 2003, 1995). This phenomenon set the stage for a source sink dynamic in the Great Lakes basin in which the less contaminated inland regions of the state supply sufficient young to keep the Great Lakes coastal populations growing in spite of reproductive productivity rates which still demonstrate impairment (Bowerman et al., 2003).

The State of Michigan has maintained a count of occupied bald eagle breeding areas and their reproductive outcomes that extends back to 1961. These records, initially collected during bald eagle nestling banding efforts, document the growth in the number of active and successful breeding areas. As nest numbers grew, accurate nest and outcome assessments could no longer be conducted by visiting every nest within the state during the breeding period and in 1977 U.S. Fish and Wildlife Service began flights for aerial nest and nestling enumeration. Arial enumeration has continued on a yearly basis, though it is now carried out through the Michigan Departments of Natural Resources and Environmental Quality. While the number of active territories has increased from 72 to over 600, outcomes are still obtained for every known occupied breeding area and 100-200 nests are visited by ground crews, which serves to verify the flight based assessments of outcome. The longevity and thorough nature of the data collection effort in Michigan have made it an extremely powerful information source.

Problem Statement

It is common practice in environmental monitoring to use summary statistics, such as means, in modeling trends, often referred to as 2-stage or derived data analysis.

Note: In keeping with established terminology, we will refer to summary statistics here as derived data, and to the method of analyzing derived data as 2-stage analysis. Likewise, we will refer to analyses conducted on raw data as 1-stage analysis.

The raw data are often not normally distributed (for example: counts, presence/absence), but 2-stage analyses use methods that assume a normal distribution under the assumption that sample sizes used will sufficiently normalize the derived data being modeled. While the Central Limit Theorem suggests that for large samples (>30, generally), what constitutes sufficiently 'large' may vary based on the underlying complexities of the data source.

When derived data are not used, there are additional complexities. For 1-stage analyses, data often consist of correlated (clustered) observations, such as repeated measurements made on the same site. This violates the assumption of independence of observations that is fundamental to parametric inferential analysis. In 2-stage analysis, this assumption is ignored by circumnavigation, and in 1-stage analysis, no adjustment can be made for this underlying correlation structure with ordinary linear regression techniques. Violating either or both of these assumptions of normality and independence can undermine the validity of significance tests.

It is the goal of this study to investigate the differences in inference that would be drawn from 2-stage and 1-stage analysis. These comparisons will allow us to assess the impacts of specifying the underlying distribution of outcome measures and accounting for correlation amongst clustered observations. It is the further hope of this assessment to

inform best practices in analysis of such data with examples of analysis from real data, at several scales.

Objectives

Here, we present a retrospective analysis of bald eagle outcome data collected as part of the efforts of the Bald Eagle Biosentinel Program (BEBP). This regional trend analysis was conducted for two outcome metrics: Productivity and Success Rate, and at several levels, detailed below. The p-values resulting from statistical tests conducted using the 2-stage vs. 1-stage methods, at each level of analysis are provided.

Levels of Analysis

Within the inland Upper Peninsula (UP), it was of interest to determine whether the trend shown by the new sampling areas of Ottawa National Forest (ONF) and Hiawatha National Forest (HNF) together, differed from the trend shown by inland UP sampling areas outside of the ONF and HNF. Within the inland Lower Peninsula (LP), it was of interest to determine whether the trend shown by the newer Manistee, Muskegon, Au Sable Rivers Area (MMA) differed from the trend shown by inland LP sampling areas outside of the MMA. These regional trends in Productivity and Success were assessed over the 20 year period from 1994 to 2013.

Lastly, within each of the new inland sampling areas (ONF, HNF, MMA), we sought to determine if newly established nests provided redundant information regarding outcome estimates, which would suggest yearly flights for the purposes of identifying newly established nests are not necessary. To this end, we modeled trends in Productivity and Success based on the information from sites established prior to 2006 compared to the

same trends based on information from newly established nests. This comparison of trends in Productivity and Success was assessed over the 8 year period from 2006 to 2013 for the UP and the ten year period from 2004 to 2013 for the LP. This difference was due to lack of newly established sites in the UP in the first 2 years of the planned comparison period beginning in 2004.

Methods

Outcome Metrics:

Terminology used in this dissertation regarding bald eagle productivity and territories is that of Postupalsky (1974).

Productivity (Prod) has been defined as the number of fledged young per active nest:

$$Productivity = \frac{\sum_{i=1}^{n} P_i}{n}$$

where,

n = Number of active nests, and

P_i = *Number of young fledged from the* i^{th} *of n active nests.*

Because Productivity estimates represent counts, in source-method analysis they were modeled using a Poisson.

Success rate (Succ) has been defined as the proportion of active nests Producing at least one fledgling:

$$Success = \frac{\sum_{i=1}^{n} S_i}{n}$$

where,

n = Number of active nests, and

S_i *= Indicator for Success for the* i^{th} *of n active nests,*

$= 1$, *if* $P_i \geq 1$,

$= 0$, *otherwise.*

Because Success estimates represent a binary indicator variable derived from counts, in source-method analysis they were modeled using a Binomial distribution.

Distribution and Trend Analysis

Retrospective analysis of bald eagle reproductive outcome data was performed on data collected as part of the efforts of the BEBP. This regional trend analysis was conducted for two outcome metrics: Productivity and Success rate, and at several levels. At each level of analysis, we provide a summary of the fit of the summary statistics to the normal distribution. The fit of the summary statistics to the normal distribution will be assessed using the Shapiro-Wilk test. While power is low for all tests of normality in samples as small as seen here (n=20, n=10, n=8), the Shapiro-Wilk test was used to assess the fit of the summary statistics to the normal distribution due to its tendency to maintain better power in small sample sizes (Razali & Wah, 2011; Zeger et al., 1988). Because of this, it is best not to rely on significance tests alone and plots of data were visually inspected.

We also provide for comparison of detection of trends of interest, the p-values resulting from statistical tests conducted using the 2-stage vs. 1-stage analysis methods, at each level of analysis. Models for trends using summarized data were built and

hypothesis testing was conducted using the GLM procedure in SAS software version 9.3 (SAS Institute Inc., 2011), and significance was tested based on the F statistic associated with the interaction between classification and year. Models for trends using raw data were built and hypotheses were tested using the GENMOD procedure in SAS software version 9.3 (SAS Institute Inc., 2011), and significance was tested based on the Z statistic associated with the interaction between classification and year in the Generalized Estimating Equation (GEE) models. This procedure allows for marginal models with the specification of non-normal distributions for the outcome variables, which gave us the ability to model Productivity as a Poisson-distributed variable, and Success as a Binomial-distributed variable. This also allowed for GEE models to include estimation of intra-site correlations at the site level that perform like conventional auto regressive correlation, but limit the correlation to *m* steps. For regional trends, 5-time step autocorrelation structure was specified based on the best fit to the whole state data, and chosen by comparing QIC for a set of ecologically based candidate structures, as recommended by Hardin and Hilbe (2003). Analyses for both modeling methods included year as a direct continuous independent variable and region and newness variables included as classifications. Interaction terms in the models, (classification*year) provided appropriate tests of significance for differences in population-averaged Productivity and population-averaged Success trends over time, as a function of region or newness of nest site.

Tests of statistical significance were conducted with $\alpha=0.05$.

Results

Distribution Fit Tests

For distribution tests, Shapiro-Wilk tests on their own provided little evidence that the assumption of normality was violated when analyzing derived productivity data or derived success data. Table 1 displays p-values for Shapiro-Wilk tests of normality for both mean productivity and mean success data at all levels of analysis. Only the derived success data for newly established and long-standing, inland LP sites displayed statistically significant deviations from normality, as assessed by the Shapiro-Wilk test. However, plots of the distributions suggested that many of the sets of derived data were substantially non-normal. Plotted data show evidence of a tendency to deviate from the normal distribution. Figures 1 through 4 show horizontal bar charts of binned values and observed data plotted against the expectation for normal quantiles. As an example of reasonably good fit to the normal distribution Figure 1, provides a visual representation of the derived productivity data at the whole state level. An example of evidence of skewness is shown in Figure 2. An Example of evidence of leptokurtosis is shown in Figure 3. Finally, an example of evidence of platykurtosis is shown in Figure 4.

Table 1. Shows p-values for Shapiro-Wilk tests of normality for summarized productivity and success data, at several levels of analysis.

Distributions Analyzed, Organized by Comparison Level	P-values for Shapiro-Wilk Tests of Normality			
	Productivity Data		Success Data	
Whole State	0.9769		0.2268	
	New	Established	New	Established
Inland UP: New (HNF/ONF) vs. Old	0.4688	0.1393	0.2593	0.8573
Inland UP: HNF vs. ONF	0.0733	0.6687	0.8457	0.7814
HNF/ONF: Gogebic vs. Iron Counties	0.9689	0.5301	0.8115	0.9317
Inland LP: New (MMA) vs. Old	0.5727	0.9464	0.7220	0.7115
Inland UP (2006-2013): Newly Established vs. Long-standing	0.3651	0.3440	0.8253	0.9099
Inland LP (2004-2013): Newly Established vs. Long-standing	0.8488	0.0107*	0.8002	0.0175*

* Signifies statistical significance at the α=0.05 level

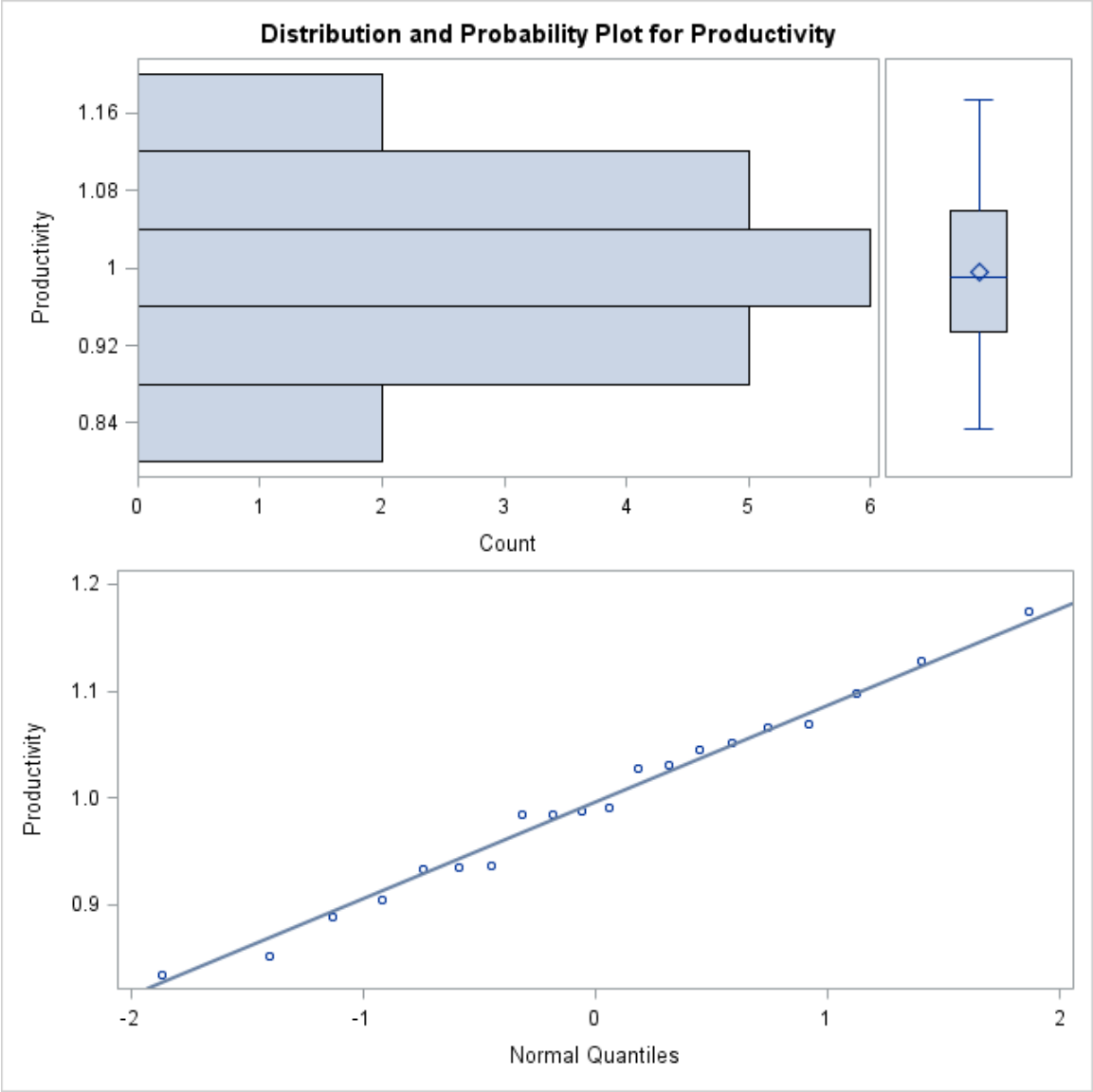

Figure 1. Distribution and fit plots for the normal distribution and derived productivity data at the whole state level, showing good conformation to the normal distribution.

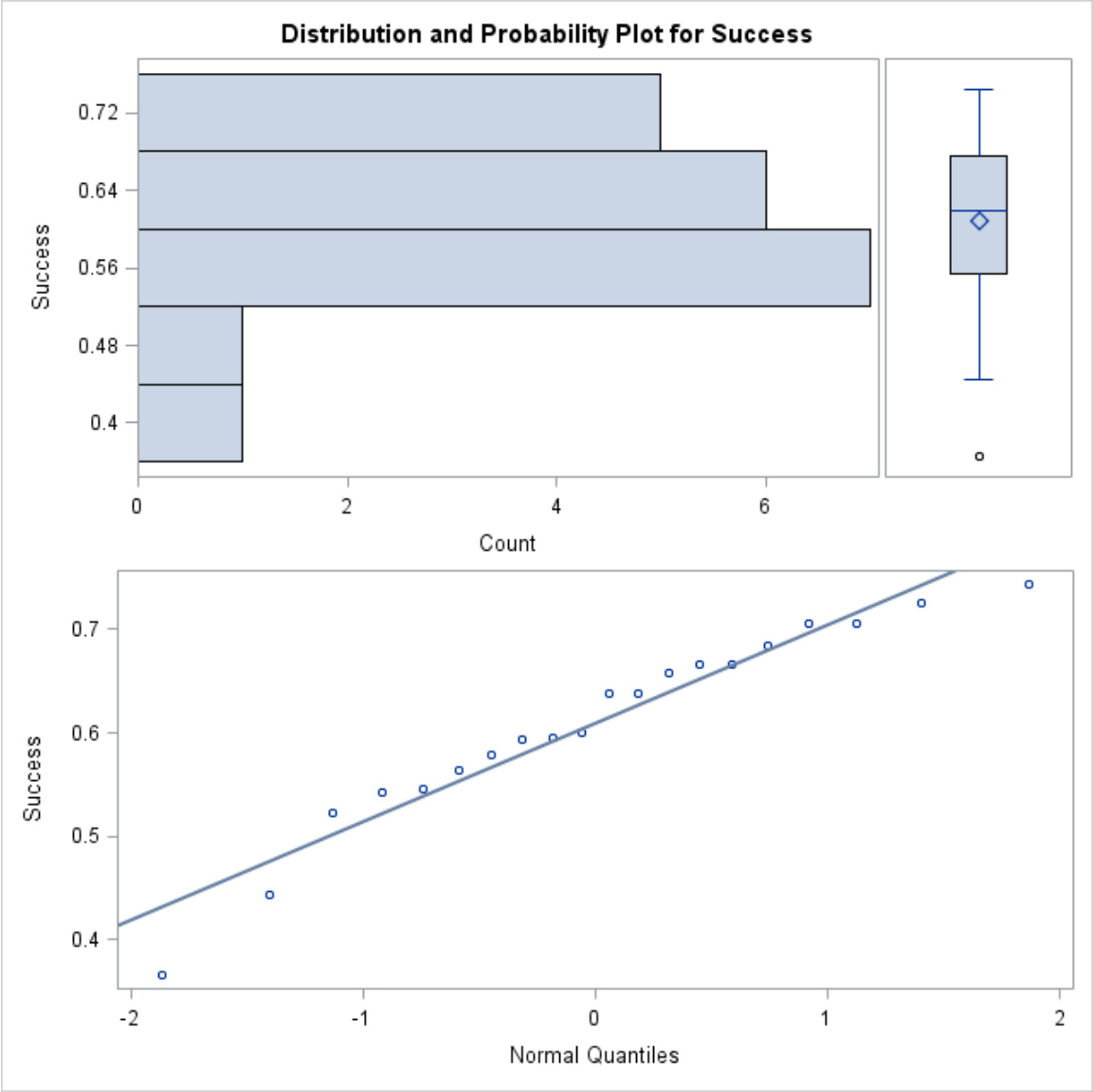

Figure 2. Distribution and fit plots for the normal distribution and derived success data for the new inland UP sites, showing evidence of skewness.

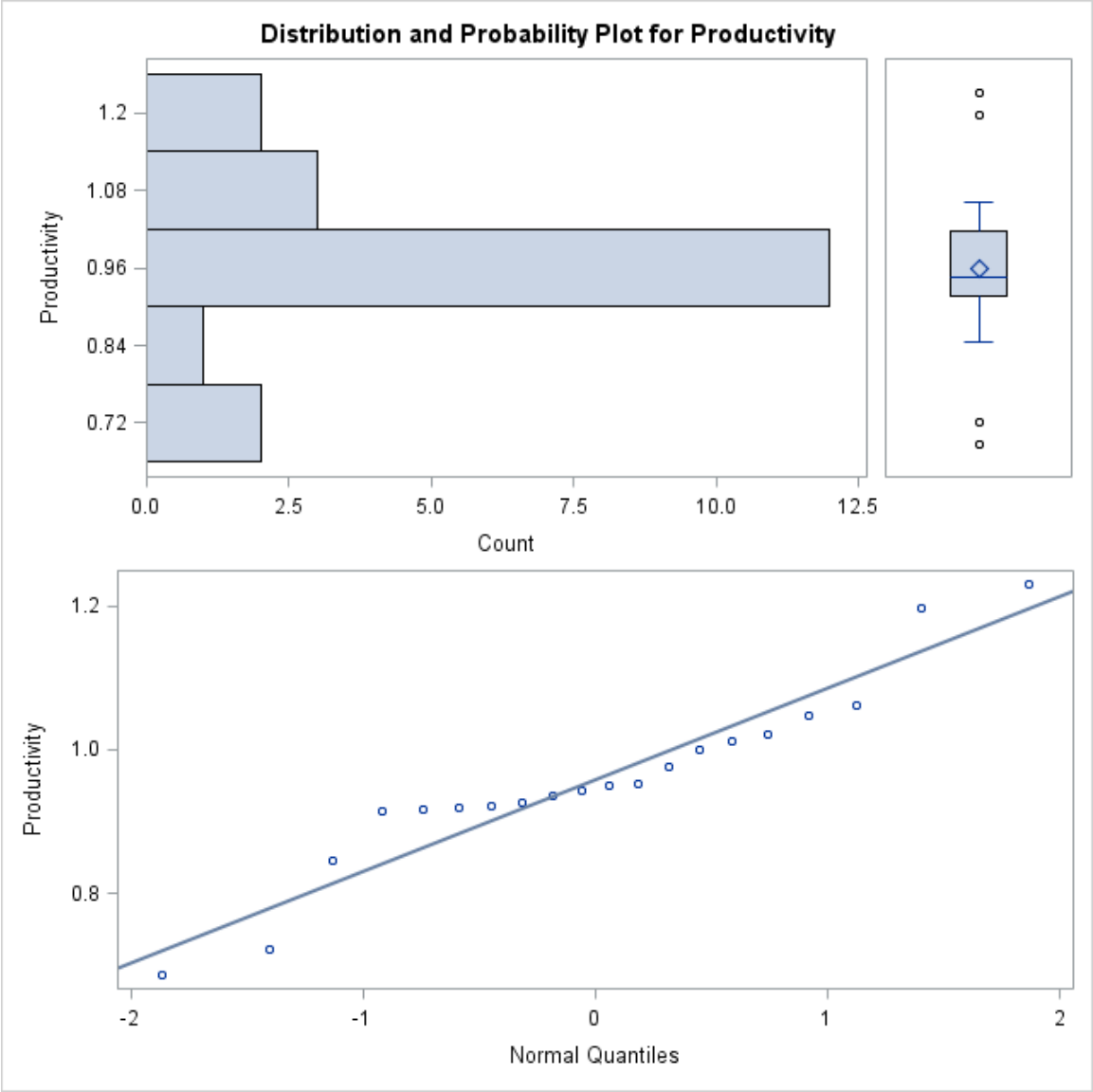

Figure 3. Distribution and fit plots for the normal distribution and derived Productivity data for the old inland UP sites, showing evidence of leptokurtosis.

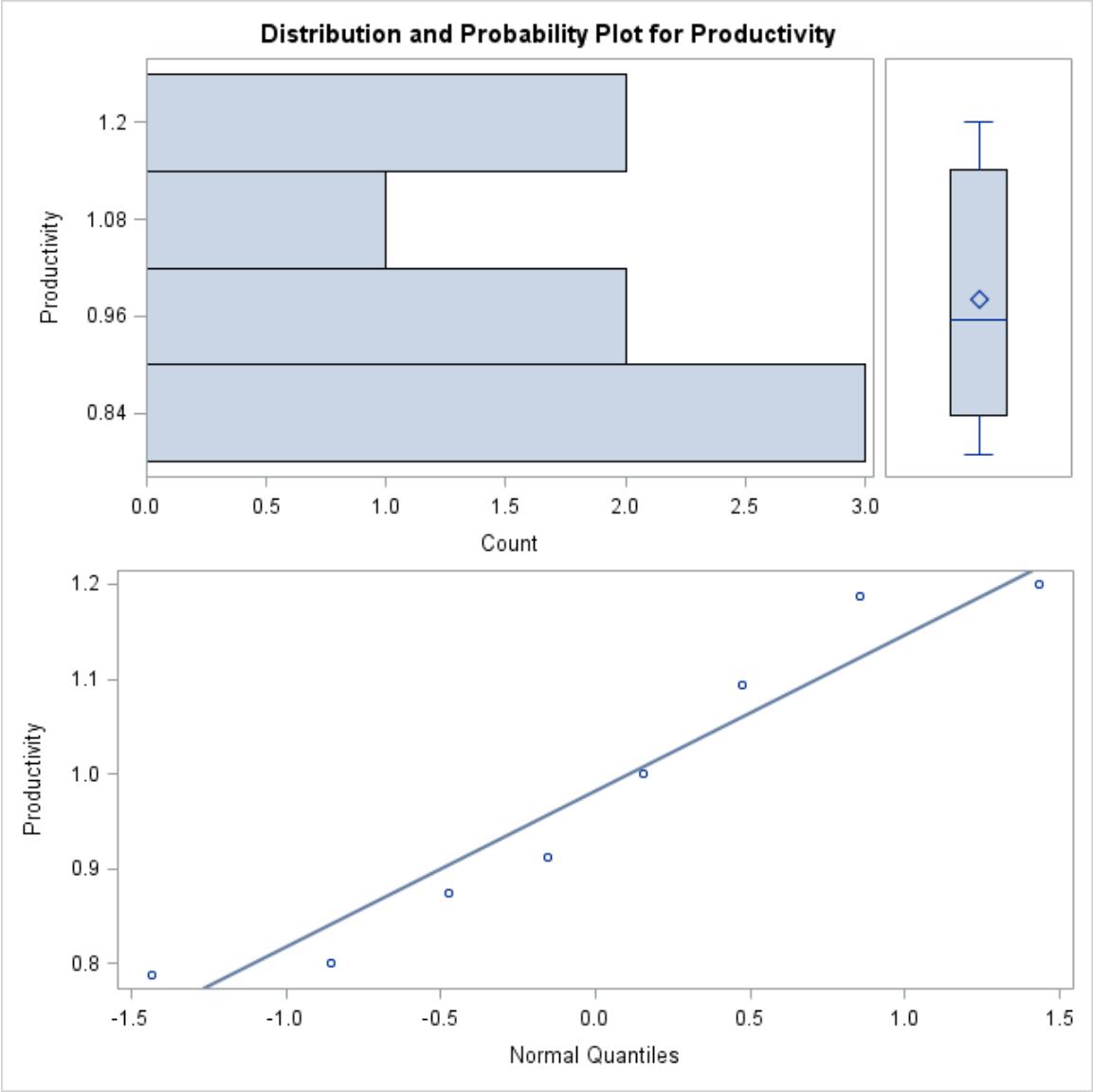

Figure 4. Distribution and fit plots for the normal distribution and derived Productivity data for the long-standing UP sites, 2006-2013, showing evidence of platykurtosis.

Trend Analysis

Analysis of bald eagle reproductive outcome data was performed on data collected as part of the efforts of the BEBP. Trends in reproductive outcome over time are of interest to the BEBP, and the goal of this analysis was to compare conclusions that would be drawn from summarized versus raw data. Analysis was conducted for two outcome metrics: Productivity and Success rate, and at several levels. Table 2 provides levels of analysis and associated p-values for significance of temporal trends in productivity from statistical tests conducted using the 2-stage vs. 1-stage analysis methods, at each level of analysis. Table 3 provides levels of analysis and associated p-values for significance of temporal trends in success from statistical tests conducted using the 2-stage vs. 1-stage analysis methods, at each level of analysis.

Table 2. Displays p-values for significant difference in productivity trends by region or degree of establishment. For comparison, p-values from the summarized data and raw data based models are shown, with levels showing instances of disagreement indicated with ([§]).

Comparison Level	**P-values for Statistical Significance of Productivity Trend by Model Type**	
	Summarized Data (Normal Dist)	*Raw Data (Poisson Dist, Clustering)*
Inland UP: New (HNF/ONF) vs. Old	0.1680	0.1543
Inland UP: HNF vs. ONF[§]	0.0832	0.0121
HNF/ONF: Gogebic vs. Iron Counties[§]	0.0122	0.0626
Inland LP: New (MMA) vs. Old	0.1026	0.0618
Inland UP: Long-standing vs. Newly Established[§]	0.0024	0.1726
Inland LP: Long-standing vs. Newly Established	0.2726	0.9408

Table 3. Displays p-values for significant difference in productivity trends by region or degree of establishment. For comparison, p-values from the summarized data and raw data based models are shown, with levels showing instances of disagreement indicated with ([§]).

Comparison Level	**P-values for Statistical Significance of Success Trend by Model Type**	
	Summarized Data (Normal Dist)	*Raw Data (Binomial Dist, Clustering)*
Inland UP: New (HNF/ONF) vs. Old	0.1328	0.1981
Inland UP: HNF vs. ONF	0.2286	0.0890
HNF/ONF: Gogebic vs. Iron Counties	0.0022	0.0238
Inland LP: New (MMA) vs. Old	0.4333	0.3607
Inland UP: Long-standing vs. Newly Established[§]	0.0058	0.1437
Inland LP: Long-standing vs. Newly Established	0.5164	0.4259

Discussion

Distribution fit tests

The good fit of the summarized data at the whole state level is not surprising, given that the sample sizes contributing to the calculated means are large (n=266-710 per year). This fit of data at this broad level to the normal distribution was supported not only by statistical tests for fit, but also by visual inspection of plotted data. This suggests that, at the whole state level, trend monitoring can be reasonably conducted using summarized data. If trends at the whole state level are relevant for monitoring purposes, then, the simpler method of using summary statistics appears sufficiently robust to violation of distribution assumptions to provide valid overall estimates. It should be noted, however, that confidence intervals on such estimates would be affected.

At finer scales of analysis, however, it seems likely that the data are not sufficiently normalized by the process of calculating means. While Shapiro-Wilk tests rarely reflected statistically significant deviation from the normal distribution for summarized data, it is important to recall the well documented low power of normal distribution fit tests in data sets with $n \leq 20$. It has been shown that while the Shapiro-Wilk test tends to perform best in smaller sample sizes, it's power can drop as low as <10%, depending on the type of non-normality in the distribution, and plots should be also inspected when assessing distribution assumptions (Farrell & Rogers-Stewart, 2006; Noughabi & Arghami, 2010; Razali & Wah, 2011; Romão et al., 2009; Seier, 2002). Plots of data suggested that, at various levels of analysis, there are likely issues of both skewness and kurtosis, both large and small. While parametric analysis is often

considered to be relatively robust minor violations of normality, the trends of interest here involve regional comparisons that quickly limit the size of samples contributing to summary statistics used, which make violations of this assumption riskier.

Trend Analysis

In addition to distribution specification, correlation structures were accounted for in modeling trends based on the raw data. Model comparisons were not conducted at the whole state level, as no inference was conducted.

It has been shown that differences occur in productivity as a function of the nature of the watershed that serves as the primary food source. Region-related influencing factors can include human encroachment, contaminant concentration, prey base, susceptibility to climate variation, and others (Bowerman et al., 1995; Buehler, 2000; Garrison, 2010; Hansen, 1987; Stalmaster, 1987; Wiemeyer et al., 1993, 1984). It is therefore of interest for the BEBP to detect regional differences in reproductive outcomes as both an indicator of the population trends for the bald eagle, and a potential indicator of ecological events.

This investigation concerned differences in average outcome trends as a function of region, and site level estimates were not of interest. For this reason, the marginal models applied using GEE were suited to our needs. In cases where inferences are being made, it is important that relevant characteristics of the data are accounted for. Distribution specification and correlation structure can substantially impact estimates of standard error. Depending on the distribution, marginal means can be incorrectly

estimated. If there is a correlation between observations, such as repeated measurements taken at the same site over several years, estimates of standard error can be over or under estimated depending on the relationship between time and classification level (Dunlop, 1994). Fortunately, when the data contain many clusters, GEE based inference is relatively robust to mis-specification of the underlying correlation matrix (Hardin & Hilbe, 2003; Stroup, 2013; Zeger et al., 1988).

There were four cases in which the conclusions of inferential analysis differed for the two model types. Three of these differences were for trends in productivity. This was for the comparison of inland UP comparing trends in the HNF and ONF, within the new inland UP areas comparing trends in Gogebic and Iron counties, and for the new inland UP areas comparing trends for newly established versus long-standing sites. For the later of these two, the 2-stage analysis model suggested significant difference in productivity trends, while the 1-stage analysis model did not. For analysis of productivity trends in the inland UP comparing the HNF to the ONF, 2-stage analysis models suggested a significant difference while 1-stage analysis did not reflect statistically significant differences. This shows that use of the raw data and full specification of the data structure leads to differences in the conclusions that might be drawn about regional trends.

The fourth observed difference was in the analysis of trend differences for success. This difference was observed for comparisons between newly established versus long-standing inland UP nests. Two-stage analysis analysis suggested that success was significantly different, while 1-stage analysis accounting for correlation within site suggested no significant differences as a function of establishment.

In the cases where differences were found, two potential complications to the 1-stage analysis existed. For the analysis of productivity trends comparing inland UP sites in the HNF and ONF and analysis of HNF/ONF sites comparing trends for Gogebic and Iron counties, the regional restrictions limited the number of clusters in the analysis drastically. This means that sensitivity to misspecification of correlation structure may have been a problem. Sensitivity to this was assessed by conducting the same analysis with the correlation specified *m*-step dependent, with *m*=1 to 4 [mdep(1)-mdep(4)], for comparison to the initially applied *m*=5 model. Table 3 shows that in none of the cases did the change in correlation structure result in a change in statistical significance of the interaction term. This provides confidence in the robustness of our conclusion of no statistically significant difference in trends for regional comparisons of HNF vs. ONF, as well as for the observed statistical significance of regional trend differences comparing Gogebic County to Iron County.

For the analyses of productivity and success trends based on degree of establishment and additional complication is apparent. As with the comparisons discussed above, the number of clusters was limited at this level of analysis, so a similar exploration of sensitivity to correlation structure specified was conducted. However, there is also a more complex relationship between time and reproductive outcome when newly established nests are analyzed. There is reason to believe that attempts at mating tend to fail more often for newly established sites than for long-standing sites, possibly due to a learning curve for the breeding pair as newly established sites are often the result of newly paired eagles, early in maturity (Best et al., 1994).

Table 4. Statistical significance with varying specifications of the correlation matrix. As initially modeled, the p-value was 0.0121 for HNF versus ONF, and 0.0626 for Gogebic vs. Iron County.

Comparison	**Correlation Structure**			
	mdep(4)	*mdep(3)*	*mdep(2)*	*mdep(1)*
HNF vs. ONF	0.0126	0.0103	0.0095	0.0125
Gogebic vs. Iron County	0.0591	0.0613	0.0662	0.0521

This means that trends over time for newly established nests are more likely to show an increasing trend for reproductive outcomes due to unstable initial attempts resolving to stable levels, rather than steady improvement from a stable starting level of productivity. Dunlop (1994) cautioned that time dependent correlations can bias estimates of significance in GEE models. For this reason, in addition to assessing the statistical significance of trend by establishment interactions with different correlation structures, it was also of interest to investigate the statistical significance of this interaction with the time series truncated to eliminate the potential influence of the learning curve.

Three of the 32 variations in specified correlation structure resulted in a change of inference where trends in productivity and success as a function of establishment were concerned. Table 4 displays the p-values for the assessment of establishment trend

interactions under three truncated models and different *m*-dependent correlation structures, with *m*=1-4. This table also makes evident the greater sensitivity to inclusion of first years of establishment, reflected by differences in p-values within the column, than to specification of correlation structure, reflected by the similarity of p-values within rows. In 2006, there was only one newly established site, so the standard error was inestimable. This, along with the addition of a second newly established site in 2007 so that the error was then estimable, are likely the driving factors behind the significance trends from 'Earliest Year: 2006' model to the 'Earliest Year: 2007' model. It further follows that as the earliest year included moves later and the significance dissipates, as you'd expect with the early reproductive attempts eliminated, thereby removing the data reflecting the observed 'learning curve' effect. Though not seen here consistently, Stroup (2013) discusses issues of 'over-modeling' compromising power, which one might expect to result in higher p-values in shorter time-series analyses with higher m-dependent structures.

If estimation is the goal, derived data are an efficient and reliable method of calculating point estimates and trends. In this analysis, no changes in inference were seen where sample sizes contributing to summary statistics were greater than 35. This suggests

that, where sample sizes are large, it may be safe to make inferences based on summarized data. It was unclear from this comparison if differences in inference regarding the bald eagle reproductive outcome regional trends would be consistently over or under estimated, but differences in inferential conclusions were evident. Due to this, and the fact that regional comparisons quickly reduce per-group sample sizes, it would be

Table 5. Statistical significance with varying specifications of the correlation matrix, with cases of changed inference indicated by (§). As initially modeled, the p-value for differences in productivity trends was 0.1726 and the p-value for differences in success trends was 0.1437.

Reproductive Outcome	Earliest Year	Correlation Structure			
		mdep(4)	*mdep(3)*	*mdep(2)*	*mdep(1)*
Productivity	**2006**	0.1696	0.1280	0.1352	0.1490
	2007	0.0611	0.0428§	0.0278§	0.0368§
	2008	0.6959	0.7394	0.7370	0.7668
	2009	0.9077	0.9612	0.9537	0.9657
Success	**2006**	0.1439	0.1459	0.1459	0.1704
	2007	0.0806	0.0822	0.0698	0.0947
	2008	0.8617	0.9080	0.8911	0.9691
	2009	0.7659	0.7915	0.7953	0.7991

good practice to use models with raw data distribution and underlying correlation specified.

The analyses presented here were conducted using generalized estimating equations, but other options are available. Site level inference would not be possible with this method, as intra-panel variation is treated as a nuisance parameter and standard errors are not estimated. Assumptions regarding missing data are stronger for GEE than for other mixed model estimation methods. These methods are frequently applied in health care setting where this is a difficult requirement to meet. This is less of a worry here, though worth considering, as there is a possibility that the same factors that result in an

inactive site would have resulted in poor reproductive outcomes. Analysis using generalized linear mixed models would be worthy of exploration, though their performance in data sets with small sample sizes has not been well explored and current functionality does not allow for autocorrelation of repeated measurements beyond standard autoregression (SAS Institute Inc., 2011; Stroup, 2013). While many have cautioned against inference based on small samples in GEE, Hubbard, et al. (2010) did not find substantial bias in estimates of standard error when simulating small sample sizes. This suggests at least, that more work is needed in this area and further comparisons between the results here and other candidate models would be informative.

Literature Cited

Best, D. A., Bowerman, W., Kubiak, T. J., Winterstein, S. R., Postupalsky, S., Shieldcastle, M., & Giesy, J. (1994). Reproductive impairment of bald eagles Haliaeetus leucocephalus along the Great Lakes shorelines of Michigan and Ohio. *Raptor Conservation Today*, 697-702.

Bowerman, W. W., Best, D. A., Giesy, J. P., Shieldcastle, M. C., Meyer, M. W., Postupalsky, S., & Sikarskie, J. G. (2003). Associations between regional differences in polychlorinated biphenyls and dichlorodiphenyldichloroethylene in blood of nestling bald eagles and reproductive productivity. *Environmental Toxicology and Chemistry, 22*(2), 371-376.

Bowerman, W. W., Giesy, J. P., Best, D. A., & Kramer, V. J. (1995). A review of factors affecting productivity of bald eagles in the Great-Lakes region – Implications for recovery. Environmental Health Perspectives, 103, 51-59.

Buehler, D. A. (2000). Bald Eagle: Haliaeetus leucocephalus. *Birds of North America*(506), 1-40.

Dunlop, D. D. (1994). Regression for Longitudinal Data: A Bridge from Least Squares Regression. *The American Statistician, 48*(4), 299-303.

Farrell, P. J., & Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation, 76*(9), 803-816.

Garrison, J. (2010). Analysis of egg-lay dates of bald eagles as a potential indicator of global climate change.

Hansen, A. J. (1987). Regulation of bald eagle reproductive rates in Southeast Alaska. Ecology, 68(5), 1387-1392.

Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. Boca Raton, Fla.: Chapman & Hall/CRC.

Noughabi, H. A., & Arghami, N. R. (2010). Monte Carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation, 81*(8), 965-972.

Postupalsky, S. (1974). *Raptor reproductive success: some problems with methods, criteria, and terminology.* Paper presented at the Conf. Raptor Conservation Techniques.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33.

Romão, X., Delgado, R., & Costa, A. (2009). An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation, 80*(5), 545-591.

SAS Institute Inc. (2011). SAS 9.3 Help and Documentation. Cary, NC: SAS Institute Inc.

Seier, E. (2002). Comparison of tests for univariate normality. *Interstat, 1*, 1-17.

Stalmaster, M. V. (1987). The bald eagle. *The bald eagle.*, i-xi, 1-227.

Stroup, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Bota Raton: CRC Press.

Wiemeyer, S. N., Bunck, C. M., & Stafford, C. J. (1993). Environmental contaminants in bald eagle eggs 1980-84 and further interpretations of relationships to productivity and shell thickness. Archives of Environmental Contamination and Toxicology, 24(2), 213-227.

Wiemeyer, S. N., Lamont, T. G., Bunck, C. M., Sindelar, C. R., Gramlich, F. J., Fraser, J. D., & Byrd, M. A. (1984). Organochlorine pesticide, polychlorobiphenyl, and mercury residues in bald eagle eggs—1969–79—and their relationships to shell thinning and reproduction. Archives of Environmental Contamination and Toxicology, 13(5), 529-549.

Wiemeyer, S. N., Lamont, T. G., Bunck, C. M., Sindelar, C. R., Gramlich, F. J., Fraser, J. D., & Byrd, M. A. (1984). Organochlorine pesticide, polychlorobiphenyl, and mercury residues in bald eagle eggs—1969–79—and their relationships to shell thinning and reproduction. *Archives of Environmental Contamination and Toxicology, 13*(5), 529-549.

Zeger, S. L., Liang, K.Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049-1060.

Conclusions

This collection of three studies addressed statistical concerns encountered in evaluating data from the Michigan Bald Eagle Biosentinel Program. Power to detect areas of concern, handling of data with censored values, and the impacts of model specification in analyzing correlated Poisson and binomial data were investigated.

Power analysis has shown that data from the Michigan Bald Eagle Biosentinel Program (BEBP) could provide a valuable resource for documenting areas of concern in the state. Log-transformed data were determined to deviate from the normal distribution to different degrees for measured PCB and p,p'DDE concentrations. Estimated standard deviations for PCB and p,p'DDE concentrations within each regional subset analyzed were close to the estimate based data from the entire state. Analysis was conducted based on state-wide standard deviation estimates of 1.23 ppb for PCB concentrations and 1.07 ppb for p,p'DDE. Necessary total sample sizes were estimated for scenarios with power of 0.80, 0.85, 0.90, and 0.95 based on changes of 10%, 15%, 20%, and 25% in observed concentration, and ranged from 74 to 1110 for p,p'DDE and 46 to 700 for PCB concentrations.

In analysis of methods of estimating central tendency, log-transformed data were determined to deviate from the normal distribution to different degrees for measured PCB and p,p'DDE concentrations, and differently for regional subgroups. Several methods of central tendency estimation were compared, both in application to the BEBP data, and in simulation studies. Based on the findings here, Kaplan-Meier (KM) statistics provide the

best estimates of geometric means in data with both left hand censorship and a right skew, like those generated by the Michigan BEBP. It may also be concluded that substitution of '½*DL' would be an acceptable treatment of censored values *only* for studies with low levels of censorship (<11% based on this analysis).

Finally, the ability to detect regional differences in trends of reproductive outcome was assessed. Regional comparisons of interest were defined based on the needs of the BEBP. Summary statistics showed varying fit to the normal distribution, and while Shapiro-Wilk tests rarely reflected statistically significant deviation from the normal distribution plots indicated that at several levels evidence of non-normality was present. Models were fit to the raw data, specifying the discrete distributions from which reproductive outcomes arise, as well as correlation structures for repeated measures. No changes in inference were seen where sample sizes contributing to summary statistics were greater than 35. Because regional comparisons quickly reduce per-group sample sizes, however, it would be good practice to use models with source data distribution and underlying correlation specified.

In summary, this collection of studies suggests that:

1. sufficient sample sizes to detect 20% and 25% increases in contaminant concentration with a power of 0.80 or 0.85 are easily obtainable and these data could help identify watersheds with emerging contaminant problems. If the area of suspected elevated contaminant concentrations was large enough that the combining of neighboring watersheds is appropriate, then greater power or smaller shifts in contaminant concentration could be detected.

2. Improved estimates of central tendency could be obtained by careful handling of data with observations that fall below the limit of detection. This is particularly important for data with levels of censorship above 11%.
3. If estimation is the goal, summarized data are an efficient and reliable method of calculating point estimates and trends. Where sample sizes are large, it may be safe to make inferences based on summarized data, but differences in inferential conclusions were evident in some cases and, it would be good practice to use models with source data distribution and underlying correlation specified.