Independent Multidisciplinary Science Team (IMST)

State of Oregon

Michael J. Harte
Robert M. Hughes
Victor Kaczynski
Nancy Molina
Carl Schreck
Clint Shock
Alan Yeakley

c/o
Oregon State University
Department of Forest Ecosystems & Society
321 Richardson Hall
Corvallis OR 97331-5752

April 27, 2009

Tom Byler
Executive Director
Oregon Watershed Enhancement Board
775 Summer St. NE Ste 360
Salem, OR 97301-1290

Dear Tom,

Enclosed is the Independent Multidisciplinary Science Team's (IMST) report titled *Issues in the Aggregation of Data to Assess Environmental Conditions*. This report is the third in a series of reports developed from the 2006 workshop on effectiveness monitoring jointly held by the Oregon Watershed Enhancement Board and the IMST.

Statewide, multiple natural resource groups and agencies collect data but the data are often not integrated into regional assessments. Data aggregation techniques could be used, in some instances, to combine disparate data sets for broader assessment of a species' or a natural resource's status or trends. This report focuses on the needs of technical staff of Oregon Plan for Salmon and Watershed partner agencies and, in particular, members of the Oregon Plan Monitoring Team. While this report discusses some of the issues related to data aggregation and a few techniques available, it is important to note that the complexities involved in such ventures require input from statisticians with experience in data aggregation.

We do not make any formal recommendations to the State in this report, but we do summarize key findings. They are:

- The potential for future aggregation should be considered in the design of data collection efforts, whether they are broad scale surveys or small research. This would include rigorous documentation of study objectives, assumptions, sampling design, variable definitions, implementation records, and database structure.
- Further use of the "Master Sample" concept [a standardized spatially balanced probability-based sampling design described in Larsen et al. (2008)[1] as a basis for investment in integrated data collection] should be considered by monitoring and research groups.
- The services of a statistician with experience in data aggregation methods should be obtained when planning data aggregation

projects. Early consultation is recommended, especially at the stages of setting objectives, evaluating studies for inclusion in the aggregation, and deciding which methods to use.

- In all analyses, uncertainty should be quantified if possible. In the use of methods (such as some models) where it is not possible, alternative conceptual frameworks and sets of assumptions, as well as model validation, should be considered.

We hope that the information in this report will be helpful OWEB and other Oregon Plan agencies. Please do not hesitate to contact us if you have any questions regarding this report.

Sincerely,

Nancy Molina
IMST Co-Chair

Carl Schreck
IMST Co-Chair

Attachment
cc with attachment:
Sue Knapp, GNRO
Sen. Jackie Dingfelder, Chair, Sen. Comm. Environ. & Nat. Res.
Rep. Brian Clem, Chair, House Comm. Ag., Nat. Res., & Rural Commun.
Rep. Ben Cannon, Chair, House Comm. Environ. & Water
Ed Bowles, ODFW
Greg Sieglitz, ODFW
IMST

[1] Larsen, DP, Olsen AR, Steven DL Jr (2008) Using a master sample to integrate stream monitoring programs. Journal of Agricultural, Biological, and Environmental Statistics. 13(3): 251–258.

Issues in the Aggregation of Data to Assess Environmental Conditions

IMST Technical Report 2009-1

Released April 27, 2009

Independent Multidisciplinary Science Team
Oregon Plan for Salmon and Watersheds
http://www.fsl.orst.edu/imst

Members:
Michael Harte
Vic Kaczynski
Carl Schreck
Alan Yeakley
Robert M. Hughes
Nancy Molina
Clint Shock

Citation
Independent Multidisciplinary Science Team. 2009. Issues in the Aggregation of Data to Assess Environmental Conditions. Technical Report 2009-1. Oregon Watershed Enhancement Board. Salem, OR.

Report Preparation
This report is based on initial drafts prepared by an IMST subcommittee consisting of Nancy Molina and Neil Christensen, with significant contributions from Kathy Maas-Hebner (IMST staff) and Dr. Don Stevens (OSU Department of Statistics). All IMST members guided the content of this report and contributed to its development. Draft outlines and content were reviewed and agreed upon by the IMST at public meetings on May 16, July 22, and December 4, 2008. The final draft was reviewed and unanimously adopted at the February 20, 2009 IMST public meeting by Michael Harte, Bob Hughes, Nancy Molina, Carl Schreck, and Clint Shock (Vic Kaczynski was absent for vote; Alan Yeakley abstained).

Acknowledgements
The IMST would like to express appreciation to the following reviewers for their insightful comments, corrections, and suggestions: Loveday Conquest (University of Washington), Steve Leider (Governor's Salmon Recovery Office, Washington), Bruce Marcot (USDA Forest Service, Mike Mulvey (Oregon Department of Environmental Quality), Tony Olsen (US Environmental Protection Agency), and Phil Rossignol (OSU Department of Fisheries and Wildlife). Additional helpful comments were provided by Susie Dunham (IMST staff).

Table of Contents

List of Acronyms and Abbreviations

CEN	Comité Européen Normalisation (European Committee for Standardization)
COSP	change of support problem
EMAP	Environmental Monitoring and Assessment Program
EPA	US Environmental Protection Agency
FGDC	Federal Geographic Data Committee
GRTS	generalized random tessellation stratified sampling design
GIS	geographic information system
IMST	Independent Multidisciplinary Science Team
MAUP	modifiable areal unit problem
NED	Northwest Environmental Data Network
NRC	National Research Council
ODFW	Oregon Department of Fish and Wildlife
Oregon Plan	Oregon Plan for Salmon and Watersheds
OWEB	Oregon Watershed Enhancement Board
PNAMP	Pacific Northwest Aquatic Monitoring Partnership
US	United States
USDA	US Department of Agriculture
USDA-FS	US Department of Agriculture – Forest Service
USGS	US Geological Survey
WSDE	Washington State Department of Ecology

Executive Summary

Sharing data across geographic and jurisdictional boundaries is one way that Pacific Northwest resource managers, policy makers, and scientists can improve their ability to make decisions about natural resources, including salmonid recovery, aquatic resource status, and watershed management. With the establishment of centralized natural resource databases and movement toward standard monitoring and sampling methods, data aggregation could be used to create regional, state-wide, or population-level assessments. Natural resource data are frequently collected in localized or spatially discontinuous patterns, and are typically gathered in surveys or studies targeted at a narrowly-focused set of questions. Inevitably, new questions arise that make it desirable to combine data sets that have different variables, or to amass data from spatially disconnected studies to address more regionalized questions. Data aggregation techniques could be used to combine disparate data sets and for 'regionalizing' data from finer to coarser scales. The goal of this report is to discuss the kinds of data that can be aggregated with suitable techniques, and the consequences of improper aggregation.

The first step in combining data is to establish objectives for the aggregation. This involves determining the extent to which available data sets can be applied to the objectives including the spatial scales being considered, the sampling designs, and the methods used in data collection. The ability to appropriately aggregate data depends on the designs of the studies under which data were collected. Because of the complexities involved, consultation with a statistician with experience in aggregation techniques is important throughout the process.

Data aggregation can manifest problems that were not present in the original studies being combined. The relationships in the data and resulting inferences can change as the level of aggregation changes. The challenge then becomes to use inference procedures that are relatively invariant to such changes, or that vary in a controllable and predictable way. The sampling designs used to collect various data sets determine how they can be combined. Ideally, a sampling design would have a built-in ability for the data to be aggregated. But many do not, and so must be retrospectively modified to allow for inclusion in an aggregation. A significant dichotomy exists with regard how inferences can be drawn from the data: whether the conclusions are based on the sampling design (generally the case in probability-based studies) or on some type of model. In nonprobability-based sampling, one must appeal to something other than the design (such as a model) to establish the connection (i.e., make inferences) between the data and the population under consideration.

Aggregation becomes more difficult when combining data from studies with different sampling designs, especially if some data were collected through nonprobability-based studies or do not completely cover the population of concern. In these situations, spatial and temporal variation cannot be assumed to have been factored into sampling in equivalent ways. The central problem becomes one of predicting data values at non-observed locations, and then performing some kind of summation over the entire population domain. Both design-based and model-based approaches exist for doing so. The difference between design-based and model-based approaches discussed in the report refers primarily to the basis upon which inferences are made and conclusions are drawn from the data, and not necessarily to the structure of the sampling design.

Model-based approaches can (and often do) combine data from probability-based and nonprobability-based designs.

Fundamentally, the statistical appropriateness of aggregating data is a function of the properties of the data as determined by the underlying sampling design. The primary need is to ensure that the relationships among variables remain constant throughout the aggregation. Depending on the nature of the relationships, such constancy may be difficult or impossible to achieve. One or more problems may be encountered during the data aggregation and analysis. Simpson's Paradox deals with problems in grouping discrete data. Parallels in data grouped across continuous spatial areas have also been recognized in geology (change of support problem), geography (modifiable areal unit problem), and sociology (ecological correlation and ecological fallacy). In addition, lurking or hidden variables and spatial autocorrelation can modify relationships between variables and confound interpretation of results.

Aggregating data from probability-based samples is relatively straightforward, and basically involves creating a single probability sample from the component studies. In order for probability samples to be combined, they must have commonality among variables of interest, and sufficient information about sampling frames and sample site selection methods to allow comparisons to be made. Methods for aggregating probability samples include combining weighted estimates, post-stratification, and direct combination into a single sample.

Combination of probability-based data with nonprobability-based data has significant limitations that must be factored into the analysis. The primary problem is that quantitative estimates of variation and uncertainty cannot be calculated from nonprobability-based data, so the validity of the results cannot be quantified. The nature and objectives of the aggregation will determine how severe a problem this may be. Methods for combining probability and nonprobability data include those that treat the nonprobability data as though it was probability-based (e.g., pseudo-random and stratified calibration approaches), models, and meta-analysis.

Once objectives and datasets for aggregation have been decided, several issues should be considered before the datasets are actually combined into a new dataset for analysis. These issues include data credibility and reliability, data inconsistencies over time and among observers, non-comparability of sampling designs and resulting data, insufficient sample sizes, differences in sampling effort, data completeness (e.g., low sampling frequency and short time-frames, and incomplete spatial and/or temporal coverage of data. Metadata records can make merging datasets together and identifying possible data incompatibilities easier. Rigorous metadata documentation includes a description of the data, the sampling design and data collection protocols, quality control procedures, preliminary data processing (e.g., derivatives or extrapolations, estimation procedures), professional judgment used, and any known anomalies or oddities of the data. Other problems that may need to be addressed include:

- data sets that are not kept electronically in their entirety (e.g., location information or date of collection may be kept on hard copies);
- data formats (e.g., metric vs. English measurements, different decimal places) and file types may be inconsistent or incompatible;

- data fields with the same name may not contain the same type of data or information (e.g., "species" may variously include common names, scientific names, or acronyms); and
- species may not be identified to the same taxonomic level (e.g., species, subspecies, or variety may not be recorded).

Based on IMST's review of the issues related to aggregating data to assess environmental conditions, the IMST makes the following observations:

- The potential for future aggregation should be considered in the design of data collection efforts, whether they are broad scale surveys or small research. This would include rigorous documentation of study objectives, assumptions, sampling design, variable definitions, implementation records, and database structure.
- Further use of the "Master Sample" concept (a standardized spatially balanced probability-based sampling design described in Larsen et *al*. [2008] as a basis for investment in integrated data collection) should be considered by monitoring and research groups.
- The services of a statistician with experience in data aggregation methods should be obtained when planning data aggregation projects. Early consultation is recommended, especially at the stages of setting objectives, evaluating studies for inclusion in the aggregation, and deciding which methods to use.
- In all analyses, uncertainty should be quantified if possible. In the use of methods (such as some models) where it is not possible, alternative conceptual frameworks and sets of assumptions, as well as model validation, should be considered.

Purpose and Scope

Sharing data across geographic and jurisdictional boundaries is one way that Pacific Northwest resource managers, policy makers, and scientists can improve their ability to make decisions about salmonid recovery, aquatic resources, and watershed management (NED 2005; PNAMP 2008). With the establishment of centralized databases such as StreamNet (StreamNet 2008) and movement toward standardization of sampling methods (e.g., Bonar & Hubert 2002; Bonar *et al.* [in press]), it is important to recognize the issues associated with data aggregation in creating regional, state-wide, or population-level assessments. The IMST realized the need to address data aggregation as it prepared a summary of the April 2006 Effectiveness Monitoring Workshop co-sponsored by the IMST and the Oregon Watershed Enhancement Board (OWEB), and a report containing scientific guidance for the use of ecological indicators to evaluate watershed restoration and salmonid recovery efforts (IMST 2006; IMST 2007). In this third monitoring report, IMST addresses scientific issues that arise when attempting to aggregate fine-scale data or data from diverse sources to answer broad-scale questions about environmental status and trends.

In this report the term *scale* [1] refers to a combination of *extent* (i.e., relative lengths, areas, and sizes including population size), and the *grain* or resolution of information (Wiens 1989). "Fine scale" is used to describe data that are collected from (or characterize processes operating within) smaller areas and/or have highly dense (or very detailed) information. "Broad scale" refers to data that describe features or processes over a large spatial extent and/or that have sparse or less-detailed information.

Regional assessment of the status and trends of natural resources often requires analysis and interpretation of data collected at different scales by multiple entities to answer a variety of questions. Natural resource data are frequently collected in localized or spatially discontinuous patterns, and are typically gathered in surveys or studies targeted at a restricted set of questions. Inevitably, new questions arise that make it desirable to combine data sets that have different variables, or to amass data from spatially disconnected studies to address more regionalized questions. The desire to wring all possible value out of available data reflects the considerable investment agencies make to acquire such data. Because it is not possible to design and implement a new study for every question that arises, methods must be used to 1) capitalize on existing data, and 2) to design future surveys to provide data that can be aggregated. In this report, both objectives are discussed.

This report is primarily intended for those engaged in the design of hierarchical monitoring frameworks or in the analysis and interpretation of natural resource data across multiple spatial scales. While this report focuses on natural resource data and assessments, data aggregation also takes place in genetics, epidemiology, medicine, as well as other disciplines (e.g., Piegorsch & Cox 1995). These and other disciplines may have data aggregation issues specific to them (e.g., in genetics, the difficulty of ensuring independence of sampled breeding populations makes

[1] In cartography, the term "scale" refers to the ratio between map length and ground distance. In this report, and in much of current ecological literature, a broader definition is used.

aggregating data from different studies problematic) or use other useful aggregation techniques not discussed here.

This report focuses particularly on the needs of technical staff of Oregon Plan for Salmon and Watersheds (Oregon Plan) partner agencies, especially the members of the Oregon Plan Monitoring Team. It describes key issues, approaches for addressing these issues, and potential problems associated with data aggregations. It assumes readers have a moderate degree of knowledge about sampling designs and statistical analysis. Readers may want to consult references cited in this report for background information or details. While this report identifies problems associated with aggregating data along with techniques to avoid or minimize them, it does not provide detailed directions for aggregating data or solutions to all problems encountered. Statisticians experienced in data aggregation should be consulted to determine which techniques are most appropriate. Finally, while this report acknowledges the existence of several ongoing programs focused on combining environmental data in statistically rigorous ways, such as the Pacific Northwest Aquatic Monitoring Partnership (PNAMP 2008), this report is not intended as a critique of such programs, nor does it propose modifications to them.

Introduction

Spatial context is a distinguishing feature of environmental data[2]. Observations are collected at specific locations and times with specific spatial resolution to answer questions at specific spatial scales. When new questions arise, as they frequently do, one can either collect new data at a scale appropriate to the new question, or use pre-existing data that were not collected specifically to answer the new question, a suboptimal solution often made necessary by limited resources.

The problem of aggregating data actually consists of two related sets of issues: one involving the combination of disparate data sets, and the other of 'regionalizing' data from finer to coarser scales, which may involve redefining or transforming variables so that they are more meaningful at the new scale. The goal of this report is to discuss the kinds of data that can be aggregated with suitable techniques, and the consequences of improper aggregation. Part 1 of this report focuses on those issues, including aspects of sample design that affect data aggregation. Part 2 presents some basic aggregation techniques for probability and non-probability data. Part 3 discusses issues related to data comparability, quality and reliability, metadata documentation, and combining datasets. A summary with key conclusions are presented in Part 4. The report concludes with a glossary, cited references, and appendices.

[2] Temporal context is also a factor in the ability to aggregate data. Combining data across time scales and different time periods has an additional set of issues (e.g., see Paine *et al.* 1998) that are beyond the scope of this report.

Part 1. Factors Affecting Data Aggregation

Key first steps in combining data are to establish objectives for the aggregation, and to ascertain the extent to which available data sets can be applied to those objectives (Olsen *et al.* 1999). The spatial scales being considered, the available data sets, the sampling designs, and the methods used in data collection all influence the ability to aggregate data. Questions to be answered in planning data aggregation include:

- What is the scale of interest for the aggregation (e.g., population, watershed, state, or region)? Can variables be defined that are relevant to the questions at this scale, and can these variables be derived from the available data? Do the finer scale data contain sufficient information about the broader scale attributes of interest, or do they tell only part of the story[3]?

- What kind of sampling designs were used to collect the data, and how do they affect the ability to make inferences? If some of the designs are not probability-based, how will the method used to select data points be factored into the aggregation?

- Will data from various studies simply be added together to create an aggregate for a larger area, or will some type of grouping or transformation be applied? If the latter, how will problems such as Simpson's Paradox (see p. 11), etc. be resolved?

- Is the geographic area across which data are to be aggregated fully covered by the studies in question, or are there spatial or temporal (e.g., seasonal or diurnal) gaps? If gaps exist, will they bias the results of the aggregation?

- What variables do the various data sets have in common? Are the definitions and measurement methods sufficiently similar that data can be aggregated? If not, what statistical tools could be used to make aggregation feasible?

- Is there spatial or temporal patterning in the environment (e.g., climate, disturbance history) that might crop up in the analysis as lurking variables, or as spatial or temporal autocorrelation?

In this section, we first discuss a fundamental issue that influences data aggregation — whether the study designs are probability-based or nonprobability-based. We then discuss problems that can develop when aggregating spatial datasets.

[3] The questions surrounding the use of attributes measured at broader scales to understand attributes at finer scales are not addressed in this report.

Sampling Design Issues that Affect Data Aggregation

Ideally, a sampling design would have a built-in ability for the data to be aggregated. But many do not, and so they must be retrospectively modified to allow for their inclusion in an aggregation. The sampling design used to collect available data is the primary determinant of how the data can be combined with data from other studies.

There are two initial questions that arise in data aggregation:

1. Are the designs *descriptive* (used to make inferences from a sample to a population; e.g., stratified random sampling by habitat types to determine fish abundance and productivity) or *analytical* (used to examine relationships among dependent and independent variables: e.g., randomized block design to examine the relationship between ground cover and planted tree seedling survival)?
2. What is the basis for choosing sample locations, i.e., are the sampling schemes probability-based or not? Descriptive designs for making inferences about populations require a sample from the same population for every population element (i.e., all locations within a sampling area have a known chance for being selected).

In contrast, analytical designs may be focused more on extreme states (or other states of interest — e.g., studying a population at sites of least and greatest abundance) in order to increase discriminatory ability. Such differences will affect how the data can be combined.

A problematic aspect of aggregation is that inferences about relationships in the data can change as the level of aggregation changes. The challenge then becomes to use inference procedures that are relatively invariant to such changes, or that vary in a controllable and predictable way. Aggregation is most straightforward with data that can be summarized with totals or averages, for example, the total number of coho salmon (*Oncorhynchus kisutch*) spawning in Oregon coastal streams. In this case, aggregation can be as simple as summing fine scale data, perhaps using weights and confidence limits that reflect the size of the spatial unit associated with the fine scale data relative to the size of the population. Data collected using probability-based sampling designs with an explicit spatial component are more readily aggregated than data collected using other designs because the necessary weights are explicit in the design.

Aggregation becomes more difficult when combining data from studies with different sampling designs, especially if some data were collected through nonprobability-based studies or do not completely cover the population of concern. In these situations, spatial and temporal variation cannot be assumed to have been factored into sampling in equivalent ways. The central problem becomes one of predicting data values at non-observed locations, and then performing some kind of summation over the entire population domain. Both design-based and model-based approaches exist for doing so. Part 2 discusses various techniques available for these more complicated data aggregations.

To produce rigorous and statistically defensible data and results, the following overarching attributes of sampling designs, compiled from several sources (Thompson *et al.* 1998; Levy & Lemeshow 1999; Thompson 2002; Roni *et al.* 2005; Larsen *et al.* 2007; US EPA 2008) are recommended:

- Precisely-stated, quantified objectives.
- Explicitly-defined target populations.
- Sample frames that accurately describe target populations.
- Sampling designs that will most efficiently provide information to meet objectives.
- Selection of sampling sites based on sampling design.
- Consistent measurement protocols.
- Statistical analyses appropriate for the sampling designs.
- Detailed documentation of all of the above.

Probability-based Sampling Designs

Probability-based samples have the characteristic that every element in the population has a known and positive (i.e., non-zero) probability of being chosen; consequently, unbiased estimates of population parameters that are linear functions of the observations (e.g., population means) can be constructed from the data. Good probability-based sampling designs also facilitate estimation of error, for example by allowing for calculation of the joint probability of including any two sample units (Levy & Lemeshow 1999, pp. 2–21). This gives the survey data a way of measuring reliability and validity that does not come with non-probability samples.

A primary advantage of a probability-based sample is that it includes a ready-made expansion factor to infer population attributes from sample attributes. Every observation has an associated weight that quantifies the proportion of the total population that is represented by that observation. This weight is defined by the sampling design, which also establishes the connection between the data and the population. In simple random sampling, for instance, the weight is the same for each data point (e.g., the population size divided by the number of samples). Knowing the weight simplifies expansion of site-specific data up to the regional level of the population because the sum of the weighted observations is guaranteed to be an unbiased estimate of the population total[4]. An estimate of the population mean is obtained by dividing the estimated total by the population size.

There are several types of probability-based sampling designs, each with pros and cons (Table 1). *Simple random sampling* is the most basic design but may be least useful for studies across large geographic areas because of possible unequal and inadequate spatial coverage of a population. Other types of designs include systematic sampling, stratified random sampling, cluster sampling, and spatially balanced sampling. Courbois *et al.* (2008) provide a discussion of how population size estimates can vary with these different types of sampling designs.

[4] This is not to imply that probability-based sampling cannot be biased. Bias may occur in probability-based sampling in execution of the sampling design, where there is an inability to sample some portion of the population, or where the sampling frame does not fit the population.

Table 1. Probability-based sampling designs and criteria for evaluating trade-offs (based on Theobald *et al.* 2007).

Criteria	Sampling design				
	Simple random	**Systematic**	**Stratified**	**Cluster**	**Spatially balanced**
Variance (inverse of accuracy of estimates)	High	High	Moderate	High	Low
Calculating the estimation of variance	Not difficult to calculate	Not difficult to calculate, but potentially biased if sampling aligns with periodicity	Moderately difficult to calculate	Moderately difficult to calculate	Moderately difficult to calculate
Spatial distribution	Poor	Good	Poor to good	Poor to good	Good
Simplicity of implementation	High	Medium	Medium	Medium	High
Flexibility for adding more sample points	High	Low	Medium	Medium	Medium

In *systematic sampling* the first point is typically randomly placed and successive points are distributed systematically (e.g., on grids or transects) throughout the sampling area. An example is the Current Vegetation Sampling Survey protocol used by the USDA Forest Service, Pacific Northwest Region, which uses a randomly started, systematic grid of primary sampling units overlaid on Forest Service lands (Max *et al.* 1996; Schreuder *et al.* 1999; Edwards *et al.* 2004). The protocol is specifically designed to allow incorporation of data from compatible surveys, such as those for lichens (Edwards *et al.* 2004). Each sampling unit contains a cluster of plots and line transects to allow the primary sampling unit to be subsampled for finer scale information (Max *et al.* 1996).

In *stratified random sampling*, samples are allocated to homogeneous subgroups (strata) that represent significant environmental differences, such as forest habitat types, and are in some way related to the population elements of concern. Each stratum is sampled independently, and the number of samples within each stratum is proportional to the size of the stratum. An alternative to stratified sampling is *unequal probability sampling* which assigns samples to target populations based on auxiliary information. The USDA Forest Service's Forest Inventory and Analysis Pacific Resources Inventory, Monitoring, and Evaluation Program (USDA-FS 1997, 1998, 2007) uses a combination of modified stratified and systematic sampling to enhance the precision of its estimates of forest health and timber volume, through a combination of photogrammetric classification and ground-based data collection.

Cluster sampling[5] is used where heterogeneous spatial groupings of population units occur. Whereas strata are internally homogeneous, clusters are like small representations of the full heterogeneity of the population, scattered across the landscape. Cluster sampling may be used for populations with patchy distributions or with uncommon species. Several modifications of cluster sampling exist. Dorr *et al.* (2008) used stratified cluster sampling in aerial surveys of cormorants (*Phalacrocorax auritis*; who tend to occur in groups) to estimate and monitor patchily-distributed bird abundance around aquaculture ponds in Mississippi. Noon *et al.* (2006) used adaptive cluster sampling (where neighboring units are added to a sample when the target sample element is found) to estimate species composition and density of terrestrial reptiles and amphibians in a tropical rainforest. Philippi (2005) also used adaptive cluster sampling to estimate abundance of a rare plant species in Florida.

Spatially balanced survey designs attempt to mimic the spatial pattern of a population (Theobald *et al.* 2007) by matching the distribution of sample points with various gradients or patterns (both spatial and temporal) observed in the population. Spatially balanced survey designs have several variations for specific needs. One of the best known is the US EPA Environmental Monitoring and Assessment Program's (EMAP) Generalized Random Tessellation Stratified Design (GRTS; Stevens & Olsen 1999; Herlihy *et al.* 2000) for monitoring the status and trends of lake, stream, river, coastal, and wetland ecosystems in the conterminous US (Stoddard *et al.* 2005, 2008; Shapiro *et al.* 2008; US EPA 2006). The Oregon Department of Fish and Wildlife (ODFW) uses the EMAP's protocol and rotating panel concept for sampling adult coho spawners, juvenile salmon, and physical habitat (Stevens 2002).

Larsen *et al.* (2008) described the use of a GRTS-based "Master Sample" approach for stream networks in Oregon and Washington. This scheme establishes a framework of potential sampling sites (points, linear networks [such as streams], or polygons) that can be sampled at a variety of spatial scales, in such a way that spatial balance relative to the resource or feature under consideration is maintained, and the advantages of a probability design are retained as successive samples are drawn (Stevens & Olsen 2004). The Oregon Master Sample consists of almost 180,000 stream sites and is currently being used in selected watersheds. The Washington State Department of Ecology has adopted the concept for use in stream sampling by several different state agencies as part of status and trends monitoring (WSDE 2006). More widespread use of the Master Sample concept could significantly strengthen the statistical rigor of stream-related data aggregation efforts in the Pacific Northwest.

Nonprobability-based Sampling Designs

Unlike probability-based samples, *nonprobability-based* samples are selected subjectively, and every element of the population does not have a known, nonzero probability of being chosen. Consequently parameter estimates calculated from nonprobability samples may be biased and any parameter estimates can only be applied to individuals from the sample (Levy & Lemeshow 1999, p. 12) with a known level of confidence. There are situations where nonprobability sampling (see Schreuder *et al.* [2001] for discussion) is useful for providing environmental data when probability sampling is not practical such as:

[5] Cluster sampling should not be confused with cluster analysis, a multivariate statistical technique.

- in studying habitats and locations of rare or cryptic species (Edwards *et al.* 2004),
- in some environmental gradient studies (Davies & Jackson 2006),
- when impediments such as expense or time constraints are present, or
- if decisions are needed immediately (Schreuder *et al.* 2001).

Nonprobability sampling can also be useful for initial screening or classification of potential sample sites, pilot studies, method development and comparison studies, and for developing testable hypotheses.

Nonprobability designs can generate useful information (Ojeda & Sahai 2002) but they are much more complicated to use in data aggregation, primarily because there can be so many permutations of design considerations. The main limitation of data collected via nonprobability-based designs is that there is no assurance they are representative of the population of interest and no ability to quantify that uncertainty. Ojeda & Sahai (2002) and Smith (1983) described methods for dealing with potential bias in nonprobability data and even described certain conditions under which non-randomness can be ignored. Bias in nonprobability sampling can result from a variety of factors, including limiting studies to variables that are believed to be important and focusing sampling on accessible areas. Nonprobability designs are also more susceptible to inadvertent exclusion of negative observations (i.e., sites where the variable of interest was not found), which can bias results. It should be pointed out that nonprobability designs can and often do use various measures to reduce bias, and that the results they achieve are not necessarily unrepresentative of the population (Ojeda & Sahai 2002). The key limitation is that the variability and bias cannot be characterized quantitatively. Nonprobability sampling designs include:

- *Observational studies* where the assignment of subjects to treatment or control groups is not controlled by the investigator; e.g., *in situ* studies of organisms in their native habitat. Observational studies are usually intended to increase knowledge about relationships, often between organisms and environmental parameters.
- *Purposive (or expert) searches* where sampling is focused on sites where the variable of interest is most likely to be found, based on current knowledge; e.g., surveys for rare species. Purposive searches are an efficient way to generate new knowledge about species distribution and habitats, but cannot be reliably used to generate population estimates (Molina *et al.* 2003).
- *Opportunistic (or convenience) surveys* where the selection of sampling locations is based on ease of access or proximity. For phenomena that are abundant, opportunistic surveys can accumulate information very quickly, but considerable bias can exist if there is spatial autocorrelation among variables.
- *Gradient studies* where sampling occurs along a known or suspected environmental gradient to ascertain relationships between the gradient and the response variables of interest.

Design-based vs. Model-based Inferences

A significant dichotomy exists with regard to the basis upon which inferences are made in analyzing data — that of whether conclusions are based on the sampling design (generally the case in probability-based studies) or on some type of model. In nonprobability-based sampling, one must appeal to something other than the design to establish the connection (i.e., make inferences) between the data and the population under consideration. The most common approach is to assume a model that describes how the data are related to the population, and then use the model to guide the inference of population attributes. Frequently, the model is a statistical distribution or curve fitted to the data. Process simulation, Bayesian network and decision models are increasingly used in natural resources disciplines (Gregoire 1998; Marcot 2006; McDonald *et al.* 2007), and these and other model forms can be used to summarize data (see Part 2).

The difference between design-based and model-based approaches as discussed here refers primarily to the basis upon which inferences are made and conclusions are drawn from the data, and not necessarily to the structure of the sampling design (Gregoire 1998). Model-based approaches can (and often do) combine data from probability-based and nonprobability-based designs. The relative merits and problems of design- vs. model-based approaches are discussed comprehensively in Hansen *et al.* (1983) and Gregoire (1998) and highlighted in Table 2.

Some scientists feel a major weakness of model-based approaches is that they lack the objectivity and unbiased character of a design-based analysis, potentially yielding results that do not accurately characterize relationships in the data (Schreuder & Williams 1995). Thus, the potential for hidden bias introduces a factor of uncertainty in models that use nonprobability-based data. Model-based approaches may be most appropriately thought of as one way to develop or refine hypotheses which may be subsequently tested by probability-based sampling (Hansen *et al.* 1983). Discussion of the use of models to combine data from disparate sources is presented on page 23.

Table 2. Pros and cons of design-based vs. model-based approaches to inference-making in biological or ecological studies (based on Hansen *et al.* 1983; Gregoire 1998; Olsen *et al.* 1999; Edwards *et al.* 2004).

Approach	Pros	Cons
Design	Lends itself well to characterizing populations (e.g., organisms, sites). Tends to be more objective and unbiased, yielding estimates that are more scientifically defensible. Inferences have known variance estimates; uncertainty can be quantified. Complexity of underlying causes or distributions of attributes do not have to be known.	Probability-based sampling may be problematic (or inefficient) where measured attributes are distributed unevenly or are extremely rare. Difficult to merge information from different studies; significant data may be ignored because they cannot be statistically included. Does not typically convey information about cause/effect. Cannot be used to make predictions to non-observed populations.
Model	Data from different studies can be merged. Can make use of non-quantitative information (e.g., expert knowledge). Can be more efficient at detecting patchy or rare phenomena, if both presence and absence have been recorded. Can be used to make predictions to non-observed locations (but with unknown error).	Requires sound hypotheses of causes or distribution of attributes, including shape (e.g., linear vs. curvilinear) of relationships among variables; in absence of such understanding, uncertainty is likely high and cannot be quantified Requires making critical assumptions that drive model outcomes, and is therefore more susceptible to error from subjectivity, bias, and misinformation. Variances can be calculated; but the reliability of predictions is not easily verifiable Problems arise when different studies incorporated in the model show conflicting results

Potential Problems when Aggregating Data

Fundamentally, the statistical appropriateness of aggregating data is a function of the properties of the data as determined by the underlying sampling design. The primary need is to ensure that the relationships among variables remain constant throughout the aggregation. Depending on the nature of the relationships, such constancy may be difficult or impossible to achieve. Below, we discuss some of these difficulties in more detail. Simpson's Paradox deals with problems in grouping discrete data. Parallels in data grouped across continuous spatial areas have been

recognized in geology (change of support problem), geography (modifiable areal unit problem), and sociology (ecological correlation and ecological fallacy). In addition, lurking or hidden variables and spatial autocorrelation can modify relationships between variables and confound interpretation of results. Other issues occur with comparability of data from different studies and are discussed in Part 3. A brief discussion of these problems follows.

Problems with Grouping Discrete Data: Simpson's Paradox

Simpson's Paradox is a statistical phenomenon in which the apparent associations of variables seem to reverse when they are grouped. It is often illustrated with contingency tables reporting frequency data (e.g., Table 3) and marginal totals (Table 3; row marked "A+B"). It occurs because there can be more than one way to stratify the variables; for example, in Table 3, fish may be stratified by watershed, or as hatchery vs. wild. When linear operators such as simple summations or means are used to look at the data, no distortions occur as a result of grouping; the aggregate of mean values is the mean of aggregate values. However, non-linear operators, such as ratios or rates, do not have this characteristic; the ratio of aggregated values is mathematically and conceptually not the same as the aggregated value of the ratios, and in the Table 3 example, the stratum used for grouping makes a difference in the outcome.

The phenomenon is best understood with examples. Table 3 shows numbers of hatchery and wild salmon from a hypothetical study where both wild and hatchery juvenile salmon are counted in two watersheds. Three years later, returning adults are counted and classified as either wild or hatchery. Assuming no counting or classification error, based on the counts aggregated (summed) over watersheds (row "A+B"), the hatchery fish seem to have higher returns (46%) compared to the wild fish (24%; bottom shaded row in Table 3). However, the *rate* of return for wild fish is greater for both watersheds (right-hand shaded column in Table 3). Averaging the return rate by watershed provides a correct return rate of 30% for hatchery fish and 40% for wild fish. Calculating return rates for wild versus hatchery fish from the aggregated totals (row marked "A+B") is clearly incorrect and misleading.

The example is contrived, but it illustrates several important issues. Perhaps the most critical is that the parameter of interest should be determined before data collection. In the hypothetical example the parameter of interest is 'return rate'. The number of fish is simply an intermediate data step before calculating the return rates. Furthermore, an attribute of a process should be determined at the spatial scale at which the process operates. In the hypothetical example, the return rates differ substantially between watersheds, suggesting that the process that determines return rate is operating at the watershed scale rather than the regional scale. Thus the variable to aggregate at the regional level is return rate, not fish count.

Another example of Simpson's Paradox is illustrated by forest research conducted by Thomas & Parresol (1989). An earlier analysis of southern pine plantations had shown that recent radial growth rates had decreased when rates were compared diameter class by diameter class, implying that the wood volume growth rates of the stands were declining. However, individual tree growth rates did not typically show this trend. This led Thomas & Parresol (1989) to weight

diameter class means by the number of trees in each class, and change the measure of growth to basal area growth. They then found that overall growth rates were increasing, not declining.

Table 3. A contingency table reporting hypothetical counts of coho juveniles and adults returning three years later. Marginal means are reported in the row titled "A+B". Shaded row shows return rates counterintuitive to return rates in shaded column.

Watershed	Numbers of fish				Return rate by watershed	
	Hatchery		Wild			
	Juvenile	Adult	Juvenile	Adult	Hatchery	Wild
A	100	10	1000	200	(10/100)*100 **10%**	(200/1000)*100 **20%**
B	1000	500	100	60	(500/1000)*100 **50%**	(60/100)*100 **60%**
A+B	1100	510	1100	260	Mean Return Rate (10 +50)/2 **30%**	Mean Return Rate (20 + 60)/2 **40%**
Return rate based on A+B		(510/1100)* 100 **46%**		(260/1100)* 100 **24%**		

Allison & Goldberg (2002) observed Simpson's Paradox in a comparison of species-level versus community-level responses to arbuscular mycorrhizal fungi across a gradient of phosphorus availability. Several individual species showed a declining response to the fungi as phosphorous increased, but when species were grouped into communities, the relationship of declining response to phosphorus weakened significantly.

Problems Grouping Data in Continuous Spatial Domains

A continuous spatial domain is a contiguous area or region, for example a political jurisdiction (e.g., city, county, or state), a natural feature (e.g., lake, estuary, watershed, ecotype), or a management unit (e.g., a state forest, ranch, agricultural field; Stehman & Overton 1996). Data collected within spatial domains can be highly correlated since samples collected near or

adjacent to one another are typically more similar to one another than samples taken a distance away. This section briefly describes problems encountered when grouping data within continuous spatial domains, gleaned from a variety of scientific disciplines, including geography, geology, and sociology, and they are inter-related. The reader may find it helpful to consult discipline-specific statistical references for more explanation.

Change of Support Problem

The change of support problem (COSP), a concept from geostatistics, arises when inferences are made from spatially transformed data, i.e., observations are made at one spatial scale but the process of interest is operating at a different spatial scale (Gotway & Young 2002). "Support" refers to the geometric size (or volume), shape, and spatial orientation of the area associated with a measurement (Gotway & Young 2002, Gotway Crawford & Young 2005). Aggregation changes the underlying 2- or 3-dimensional space represented by a variable, creating a new variable with different spatial and statistical properties (Gotway & Young 2002, Gotway Crawford & Young 2005). Table 4 provides example of COSPs. In mining operations, COSPs receive considerable attention in calculation of volumes of material over large areas from core samples. Meteorological data are also subject to COSPs, where a continuum (for example of temperature or precipitation) must be inferred from point data. Gotway & Young (2002) provides an in-depth discussion on COSPs and some geostatisical solutions. Gelfand *et al.* (2001) addressed spatial and temporal aspects of COSPs in determining ozone levels over different areas of Atlanta, Georgia. Ravines *et al.* (2008) also addressed spatial and temporal aspects of COSPs in rainfall and runoff data from the Rio Grande River basin, Brazil.

Table 4. Examples of change of support problems. Table reproduced from Gotway & Young (2002) with permission. Reprinted with permission from *The Journal of the American Statistical Association*. Copyright 2002 by the American Statistical Association. All rights reserved.

We observe or analyze	But the nature of the process is	Examples
Point	Point	Point kriging; prediction of under-sampled variables
Area	Point	Ecological inference; quadrat counts
Point	Line	Contouring
Point	Area	Use of areal centroids; spatial smoothing; block kriging
Area	Area	Modifiable areal unit problem; areal interpolation; incompatible/misaligned zones
Point	Surface	Trend surface analysis; environmental monitoring; exposure assessment
Area	Surface	Remote sensing; multiresolution images; image analysis

Modifiable Areal Unit Problem

In the absence of variability, the unit of aggregation has no impact on the value of a quantity expressed as a per unit value (e.g., velocity expressed as m/sec, density as g/m^3, or biodiversity as number of species/km^2). The result is the same regardless of the size of the measurement unit. However, real systems always have some variation, so the result of aggregation can be highly influenced by the measurement unit size and the variation encompassed therein. Yule and Kendall (1950) noted that correlations between variables measured on *modifiable* units such as field plots or geographical areas depend on the size of the unit in contrast to variables measured on *non-modifiable* units such as persons, automobiles, or trees. Openshaw & Taylor (1979) describe the issue of variability in a geographical context as the *modifiable areal unit problem* (MAUP). The MAUP is a potential source of error that can affect studies that aggregate spatial data; if relationships between variables change with selection of different areal units then the reliability of the results decreases.

The MAUP arises because spatial units are modifiable (in the sense that they can be aggregated to form other units or to change in shape) and are often arbitrarily determined (Jelinski & Wu 1996)[6]. There are two components to the MAUP, the "scale (aggregation) effect" and the "zonation (grouping) effect" (Figure 1). The scale effect describes the inconsistency of statistical results from various levels of aggregation (Openshaw 1983; Amrhein 1995; Wong 1996). Aggregation decreases variances and "smoothes" the resulting values such that information is lost (Wong 1996). Smoothing applies to all variables or attributes associated with spatial observations but the amount varies with the level of aggregation (Wong 1996; e.g., compare variances in Figure 1, parts a, b, c). The zonation effect refers to the variability of statistical values when areal units vary in size and shape while the number of units remain the same (Openshaw & Taylor 1979; Openshaw 1983; Jelinski and Wu 1996, Wong 1996; e.g., compare the means and variances in Figure 1, parts d, e, f).

[6] Political units, watersheds, hydrologic units, and ecoregions are sometimes biased or inaccurate classifying units for ecological monitoring data, although they are extensively used (McCormick *et al.* 2000; Van Sickle & Hughes 2000; Waite *et al.* 2000; Omernik 2003). For example, hydrologic units have been found to be poorly related to patterns in aquatic biota and water quality and quantity in some cases. (Omernik & Bailey 1997; Griffith *et al.* 1999; Omernik 2003; Brenden *et al.* 2006; Hollenhorst *et al.* 2007)

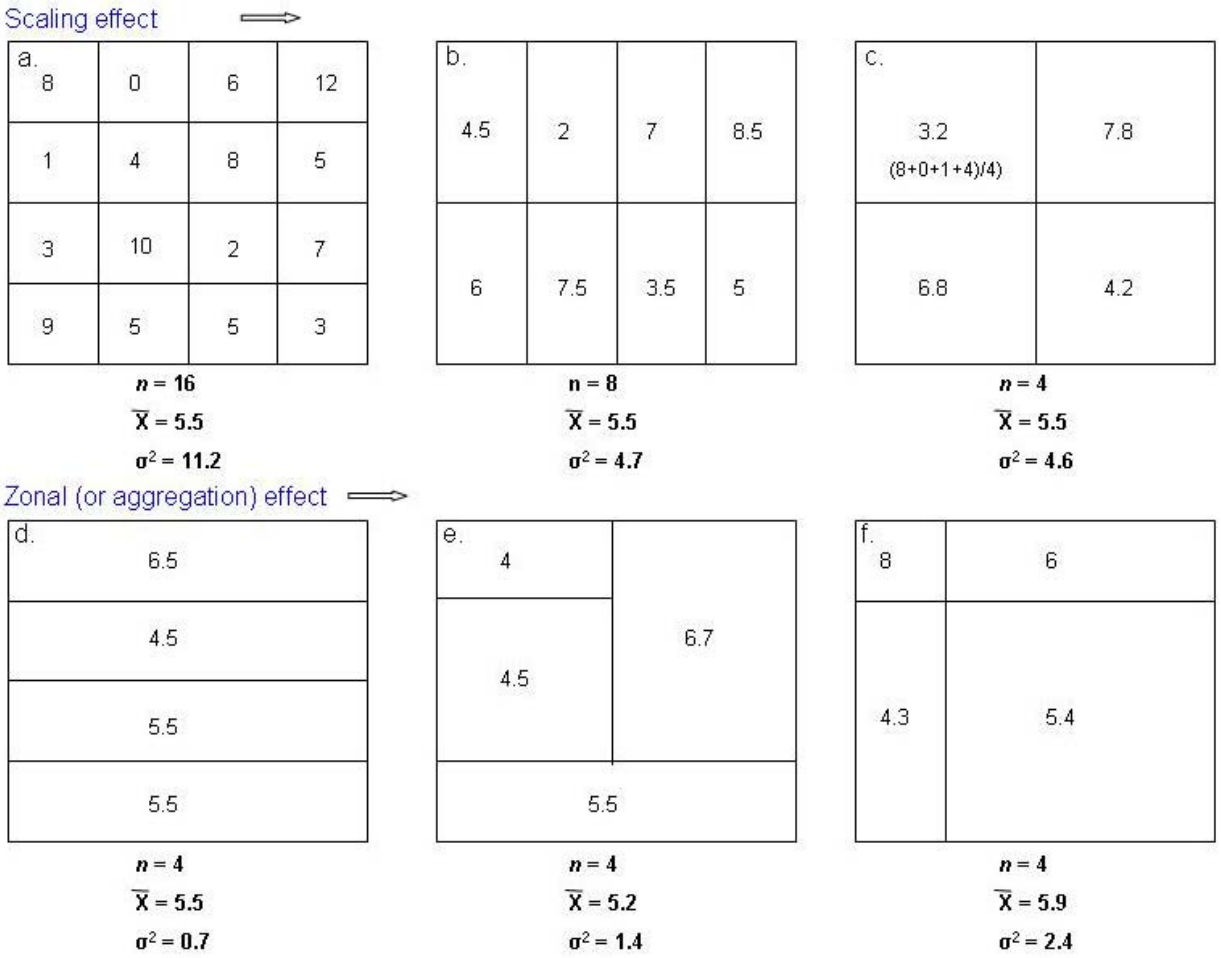

Figure 1. Hypothetical example of scale effects as smaller units are aggregated into larger units and zonal effects as spatial unit shapes or sizes are modified. (Note the way individual variables were calculated under aggregation as shown in part c.). In the scale effect (parts a, b, c), the units within each grouping are similar size; the mean value does not change but the variance declines with increased aggregation; leading to a loss of information on spatial heterogeneity (Jelinski & Wu 1996). In the zonal effect, both the mean and variance change when areal units are reconfigured. (Figure 1 is based on Amrhein 1995, Jelinski & Wu 1996; Wong 1996).

Svancara *et al.* (2002) examined how the MAUP affected the statistical relationship between elk (*Cervus elaphus*) recruitment and three independent variables (forest productivity, the proportion of non-batholith land across the summer range, and mature bull elk density) when game management units were aggregated to three different levels and three different configurations, based on either average calf:cow ratios, geographical continuity, or random groupings. Svancara *et al.* (2002) found inconsistencies in variances, correlation coefficients, regression parameters, and regression model fit (coefficient of determination) across aggregations. Differences were not only dependent upon the unit configuration and level of aggregation but also on the variable of interest (Svancara *et al.* 2002).

From a series of controlled statistical simulations, Amrhein (1995) concluded that the effects of MAUP on aggregation depend on the statistics calculated (e.g., means, variances, regression coefficients, or Pearson correlation coefficients). Based on these simulations, Amrhein concluded that the MAUP in spatial analysis does not appear to be as pervasive or unpredictable as described in earlier literature, and aggregation effects may be more easily identified and dealt with than once thought.

Ecological Correlation and Ecological Fallacy

Ecological correlation is a term originally used by sociologists to refer to correlations between variables that are group means (e.g., the correlation between salmonid/seafood consumption rates and per capita income) as opposed to individuals. Clark & Avery (1976, p.429) stated that a significant "*disadvantage of using aggregate data is the inherent difficulty of making valid multilevel inferences based on a single level of analysis*". Variables used in individual correlations (such as weight, age, or length) are descriptive of properties of individuals, while the statistical objects in an ecological correlation are properties of groups (e.g., rates, percentages, or means (Robinson 1950). Robinson concluded that ecological correlations between aggregated individual properties can be misleading. Assuming what holds true for the group also holds true for an individual is an inappropriate extrapolation or *ecological fallacy* (Johnson & Chess 2006). Similarly, what holds true for a region does not necessarily hold for an area within the region. For example, the relationship between years of schooling and support of environmental issues on a state-wide basis may be quite different from the relationship between average years of schooling and support of environmental issues on an individual basis. Ecological correlation can also occur when data are pooled across time; for example, Schooley (1994) found that black bear (*Ursus americanus*) habitat selection varied by year but was similar between two study areas in individual years. However, when data from individual years were aggregated, selection at the two sites appeared to differ. In this case, the annual variation was lost in the aggregation, leading to incorrect inferences about selection between the two sites.

Other Problems

Pseudoreplication

Generally the more replicates (sites at which independent applications of the same treatment occur) that are used, the greater the statistical precision of the resulting data analysis. Lack of

sample independence can lead to *pseudoreplication,* a common error in ecological studies (Hurlbert 1984; Heffner *et al.* 1996; Death 1999; Millar & Anderson 2004). When samples are pseudoreplicated, the natural random variation exhibited by a variable is not properly quantified (Miller & Anderson 2004). For example, repeated sampling of fish abundance from the same stream reach does not reflect the variation inherent in the stream system as a whole. Randomly drawing samples from different stream reaches experiencing the same stressor intensity or treatment would allow more accurate estimation of variability in the fish abundance response, although there is debate in the literature about whether this truly eliminates pseudoreplication (McGarvey & Hughes 2008).

Pseudoreplicated samples appear larger in size than they truly are, giving the illusion of statistical power where little exists. Consequently, inferential statistics must be used with great care because most tests are designed for samples of independent observations. Inaccuracies are typically manifested in biased standard errors that misrepresent (typically underestimate) variation in the data and artificially inflate the significance of statistical comparisons. Pseudoreplication greatly increases the chance of reaching conclusions of significance for phenomena that only happened by random chance.

Spatial Autocorrelation

Spatial autocorrelation occurs when measurements taken at sites in close proximity exhibit values more similar than would be expected if variation were distributed randomly across space or through time. In other words, the value of a measurement depends on, or can be predicted from values measured at nearby sites. Spatial autocorrelation can result from characteristics inherent to a species growth or ecology (e.g., clonal growth, conspecific attraction) or external factors (e.g., the tendency for disturbances to be correlated with vegetation patterns; Lichstein *et al.* 2002). Spatial autocorrelation is particularly problematic in model-based approaches (Marcot, pers. comm.[7]), and aggregation efforts that entail the use of spatial models should take this into account. Fortin *et al.* (1989) and Lichstein *et al.* (2002) describe methods for identifying and overcoming spatial autocorrelation in ecological analyses.

Cross-Scale Correlation

Researchers pursuing multi-scale studies of habitat relationships have documented cross-scale correlation, i.e., correlations between habitat variables measured at different spatial scales (Battin & Lawler 2006). Where cross-scale correlations exist, erroneous conclusions may be drawn about strength of relationships among predictor and response variables measured at a particular spatial scale (Battin & Lawler 2006; Lawler & Edwards 2006). For example, the presence of large, old conifers may be a good predictor of the occurrence of a species at a fine (e.g., stand-level) scale. However, the 'tree size/age' predictor variable might correlate with broader scale variables, such as average stand age, that are not necessarily good predictors of the species' occurrence. In this example, the presence of remnant old conifers might be masked at the broader scale by the inclusion of more abundant small young trees in the stand age mean. It would thus be erroneous to conclude that stand age predicts the species' presence simply because stand age correlates with tree size. It is therefore essential, in designing data aggregation, to investigate the

[7] Bruce Marcot, USDA Forest Service, Portland, OR. pers. comm. August 3, 2008.

possibility that observed relationships are the result of a process actually operating at a scale finer or coarser than the scale at which the analysis was conducted. Battin & Lawler (2006) reviewed statistical techniques for detecting cross-scale correlations among variables measured at different spatial scales. Lawler & Edwards (2006) demonstrated the use of variance decomposition (Whittaker 1984) as a diagnostic tool for revealing the amount of variation in a variable of interest explained by habitat variables measured at different spatial scales.

Lurking Variables

Bringing data together from various sources across extensive spatial domains may mask the influence of unknown variables. The association between two variables *X* and *Y* can be induced or modified by the presence of a third *lurking* variable (or covariate, also called a latent variable) that has not been identified. In some cases, it may be extremely difficult to identify a single variable or small collection of variables that act as modifiers. Spatial pattern and abiotic conditions can be common lurking variables that account for variation in environmental and ecological data. In aquatic assemblage data, the size of the water body and geographic location from which samples are drawn can have enormous implications for results (Hughes & Peck 2008), and calibration techniques are available (e.g., Fausch *et al.* 1994; McGarvey & Hughes 2008). Year-to-year and seasonal variability may also confound aggregation results. The lurking variable can be extremely critical if, for example, *X* represents a management strategy or action, and *Y* represents the resulting environmental condition. In this case the lurking variable problem potentially confounds attempts to correctly ascertain the effects of management actions, unless adequately accounted for in the sampling design and data analysis.

Studies can be designed to enhance recognition of lurking variables. One technique is to ensure that the same levels of the controllable treatment, or predictor variables, are applied over all watersheds or geographic regions. Although there may still be a hidden geographic effect on the response, the presence of a lurking variable may be easier to discern with uniform treatment levels. Cressie (1996) describes an approach for adjusting a regression by explicit inclusion of a lurking geographic variable (see Appendix A).

Part 2. Aggregation Techniques

Typically, information about the condition of lands and natural resources can be developed through a variety of methods that differ significantly in sampling design. Important sources of information often include localized observational studies and randomized-treatment experiments, aerial photography or remote sensing, and probability and/or nonprobability-based sampling efforts at various spatial scales (McDonald *et al.* 2007). The task of weaving disparate pieces of information into a scientifically-defensible assessment can be analytically daunting.

As discussed in the previous section, the ability to appropriately aggregate data is highly dependent on the sampling designs used. The data can be either random or nonrandom. They may be informative or not with respect to the variable of interest. They may include bias and be subject to measurement error; and they may have a temporal or longitudinal data structure combined with a spatial structure. All these factors may affect population estimates and inferences (Schreuder *et al.* 2001).

This section describes techniques for aggregating data from both probability and nonprobability-based sampling designs. Since the statistical methods needed to combine environmental information from different studies will often require case-specific formulations (Cox & Piegorsch 1994), this section should not be viewed as an all-inclusive summary of techniques but rather is designed to demonstrate some of the steps that can be taken to accomplish data aggregation. Olsen *et al.* (1999) cautioned that if studies were designed without the anticipation of combining additional data, some of the approaches described in this section may not be feasible.

Combining Data from Different Probability-based Sampling Designs

For reasons discussed in Part 1 of this paper, combining data from different studies is easiest if the sampling designs are probability-based (Olsen *et al.* 1999). To be combined, datasets from probability-based studies must have variables in common (or variables that can be transformed to achieve commonality) and must be capable of being restructured as a single probability sample (Larsen *et al.* 2007). Cox & Piegorsch (1994, 1996) describe the following methods for combining data from two or more probability-based surveys.

The first method combines weighted estimates from separate probability-based samples. The estimates for the parameter of interest and its variance are computed for each probability-based sample, then each estimate is weighted inversely proportional to its estimated variance, and then the weighted estimates are added (Cox and Piegorsch 1994, 1996). This results in *"a design-based unbiased minimum variance combined estimate"* (Cox & Piegorsch 1996, p. 300).

A second method is based on post-stratification (Cox & Piegorsch 1996; Olsen *et al.* 1999). Strata are defined by using shared frame attributes or subsamples that partition the two probability-based samples. Both samples are post-stratified by revising sample unit weights proportional to the new stratum size. Revised estimates are then computed for the parameter(s)

of interest. Cox & Piegorsch (1996) indicated that dual-frame estimation can be used to combine the estimates or to estimate a non-frame variable or an index based on frame variables.

In the third method, two probability-based samples are directly combined into one probability-based sample (Cox & Piegorsch 1996). The probabilities of each sampling unit's inclusion in the combined sample are computed from their first- and second-order inclusion probabilities in the original samples.

Larsen *et al.* (2007) combined stream monitoring data from two probability surveys implemented in Oregon to demonstrate how incorporation of design principles can facilitate data aggregation. Their approach also illustrates elements of the methods described by Cox & Piegorsch (1996). The sources of data were the ODFW integrated monitoring program for salmonid populations, stream habitats, water quality and aquatic biotic assemblages (Nicholas 1997), and the USDA-FS Aquatic and Riparian Effectiveness Monitoring Program, which was focused on indices of watershed health (Reeves *et al.* 2004). Even though these two efforts were targeted at questions at different spatial scales and different indicators, the data could be easily combined because of adherence to sound design principles, and because the details of the survey frames and sample selection methods were well-documented, allowing the surveys to be aggregated into a single probability sample.

The national Wadeable Stream Assessment (Olsen & Peck 2008; Paulsen *et al.* 2008) provides another example of how sampling designs from large regional surveys of aquatic biota can be combined into a single evaluation, in this case of all wadeable streams for the conterminous US. Both surveys included in the Wadeable Stream Assessment used the US EPA's River Reach File as the basis of the sampling frame. The Wadeable Stream Assessment initially selected a subset of perennial streams in the EMAP's western stream survey, which used a stratified, unequal probability design. These were then combined with a new unequal probability survey design on the remaining conterminous states to assess all wadeable streams and rivers in the lower 48 states.

Combining Data from Probability-based and Nonprobability-based Sampling Designs

Many environmental monitoring programs acquire data via both probability and non-probability methods. Consequently, the need arises to pool data from both types of sampling into a single analysis; some approaches for doing so are discussed here.

There are serious caveats in applying any of the methods presented in this section. There should be some strong evidence that the samples used are indeed synoptic (i.e., sampled from the entire population), or at the very least, lack of any evidence suggesting preferential selection. Also, the interpretation of any variance estimate is open to question, and in some methods quantifying uncertainty is problematic.

Analysis outcomes must be understood within the context of the underlying conceptual framework and assumptions, and alternative models should be considered. For example, ODFW once used nonrandom index sites, that had been selected previously because they were highly

productive, to estimate total escapement of Oregon coastal natural coho salmon (*Oncorhynchus kisutch*). Those sites, the assumptions that led to their use and the models developed to estimate total returns led to 3 to 5 fold overestimates of coho numbers and decades of unsustainable commercial harvest limits. After a probability sample was used and those earlier overestimates were recognized, commercial harvest of coho was significantly reduced (Jacob & Cooney 1995; Paulsen *et al.* 1998; Hughes *et al.* 2000). It is not just the biological sciences that are susceptible to such challenges. Writers in such diverse disciplines as philosophy of science (Kuhn 1970), economics (Daly 1973), ethics (Nash 1989), and geography (Diamond 2005) have cautioned against assuming the correctness or truthfulness of current empirical models and the conceptual frameworks upon which they are based, regardless of the subject area. Recent experience with financial models, in particular, should remind us to critically examine all model assumptions, coefficients, and their conceptual frameworks (e.g., Hendry & Ericksson 2001; Perkins 2004; Greenspan 2008).

Non-probability data may be placed into a probability-based sample context by simply treating the non-probability data as an instructed, simple random sample. While this solution is suboptimal in a rigorous statistical sense, it allows a "pseudo-probability" structure to be created for the sample. To do so, some information about the entire population must be available, in addition to information from the sample. If nothing else, the locations of samples and the range of the population are usually known. A rich array of remotely-sensed information from satellites and aerial photography may also be available to give context to the sample. The pseudo-probability approach is used in methods proposed by Overton *et al.* (1993), and Brus & de Gruijter (2003), described below. There may be other approaches being developed that have not yet been discussed in the scientific literature.

If only the locations of the sample sites are known, there is still some recourse. All environmental populations have spatial structure because locations near one another are subject to the same natural and anthropogenic stressors and influences (i.e., they are statistically autocorrelated). One approach to imputing a pseudo-probability is simply to say that a sample site represents all of those population elements closer to that site than to any other sample site. The size (number, length, area, or volume) of the total of those elements is then used as a weight for that sample point.

Pseudo-random and Stratified Calibration Approaches

In this set of approaches, the first step to combining data is to choose valid probability-based samples and "found" data sets. Found sites are chosen from the overall non-probability sample that conforms to the probability sample characteristics (Overton *et al.* 1993). One of two methods can be used to determine similarity between probability and non-probability samples and to produce population estimates: *pseudo-random* and *stratified calibration* (Figure 2). The pseudo-random approach is used when the variable of interest from the found database was also measured in the probability-based survey. If the variable of interest is only known for the found data, then stratified calibration is used.

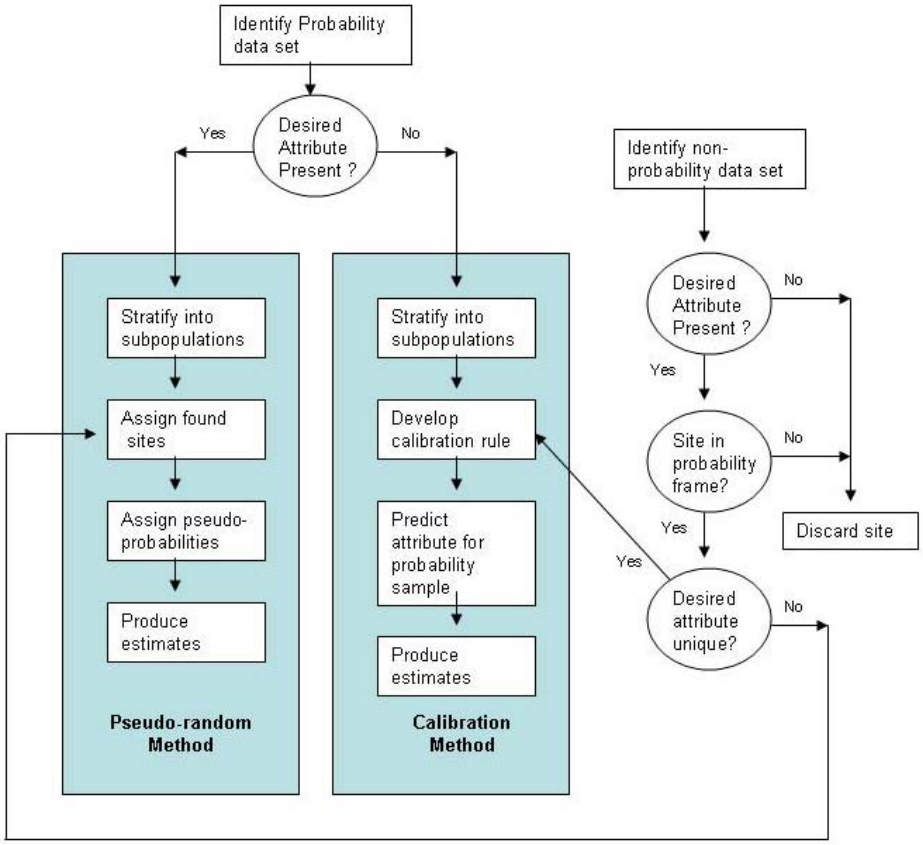

Figure 2. A schematic of Overton *et al.*'s (1993) process for combining non-probability based data with a probability-based data set. Figure redrawn from Overton *et al.* (1993) with kind permission from Springer Science and Business Media (copyright 1993).

The concept behind the pseudo-random approach is closely related to the post-stratification technique that is sometime used to improve a poorly randomized sample after-the-fact. To combine the samples, the sampling frame attributes are used to classify the probability-based sample into homogeneous groups or subpopulations (Overton *et al.* 1993). Found sites are then assigned to the subpopulations. Pseudo-random samples are defined by treating the non-random sample as if it were a stratified random design with simple random sampling within the strata. Population estimates can then be calculated from the combined data.

The stratified calibration technique is used when the desired population attribute was not measured in the probability sample. The initial steps are the same as for the pseudo-random approach described in the preceding paragraph; similarity between the data sets is established, the probability sample is stratified, subpopulations are identified, found sites are assigned to the subpopulations, and predictor equations for desired attribute are developed for each subpopulation (Overton *et al.* 1993). If two subpopulations have similar predictor relationships they are combined, if not they are kept separate. Some populations may not have corresponding found data so no predictor equation can be developed. The desired attribute is then predicted for the probability sample and population estimates can be calculated.

Astin (2006) combined nonprobability-based data (targeted/judgment sampling) with probability-based data and census data to select and calibrate water quality indicators in the Potomac River basin. Data were from Maryland, Virginia, and Pennsylvania. Because each monitoring group used variations of the US EPA's *Rapid Bioassessment Protocols for Streams and Rivers* (Plafkin *et al.* 1989) an additive or multimetric framework based on the protocols was used to combine the data. Astin (2006) assumed that repeat observations taken at fixed sites were independent and that the sites were representative of the range of conditions found in the Potomac River basin.

Overton *et al.* (1993) cautioned that because found sites were not chosen randomly, there is an unprovable assumption that the sites are representative. Brus & de Gruijter (2003) offer that if one is not confident in the representativeness of the found sites or one does not want to make this assumption then the methods proposed by Overton *et al.* (1993) should not be used. This led Brus & de Guijter to develop an alternative method.

Brus and de Guijter Approach

The Brus & de Guijter (2003) approach is a relatively new method proposed to combine probability and non-probability data which maintains the assumption of representativeness. The validity of results from estimating means of the non-probability data is ensured by collecting and combining additional data through probability-based sampling. Brus & de Guijter's approach involves overlaying a grid onto the non-probability data and randomly sampled points and calculating the difference in the means of the probability and non-probability data by interpolation through point-kriging. The error in estimating the mean for each non-probability sample is calculated by the difference of the true mean and the average of the kriged values: this error is then used to calculate measures of bias and variance. Brus & de Guijter (2003) consider the resulting estimators to be fairly unbiased, even when the non-probability sample is very biased.

Models

Another avenue for combining probability and non-probability data is to side-step the statistical differences and apply a model-based analysis to the pooled data. The validity of the resulting inference then rests entirely on the assumed model (see earlier discussion of model-based approaches, page 9).

Using a model-based approach involves constructing a model from the available information regardless of sampling design, and then using a test data set (that has not been used in the construction of the model) to validate the model, being sure to include all subpopulations of the target population in the sample. In such an approach, the model is viewed as a hypothesis about the relationships among variables.

Models can be spatially explicit or not. The Mid-Atlantic Integrated Assessment (US EPA 2002) provides an example of a spatial model that combines probability and non-probability data. The agency used a regression approach to combine probability-based stream sample data from a variety of sources (Stoddard *et al.* 2006) with landscape metrics, using GIS techniques to expand

estimates of water quality parameters to spatially continuous surfaces (i.e., depictions of actual landscapes), including non-sampled areas (e.g., Jones *et al.* 2001b, 2005). The model was used to examine a variety of other questions, including breeding bird community integrity and land use change (O'Connell *et al.* 2000; Jones *et al.* 2000, 2001a). A very similar approach was used to model the distribution of fish species in coastal watersheds of Oregon and Washington (Herger *et al.* 2003).

NetMap (Benda *et al.* 2007) is another example of the use of a spatial model to combine information from multiple sources. Using a digital terrain database, NetMap generates a host of topographic attributes that, in combination with other data layers (e.g., climate information, forest stand age, or road density) and research studies, are used to calculate and map indices of erosion risk, habitat suitability, sediment and wood supply, etc. within the context of stream networks.

Bayesian belief network models (Ellison 1996) have been used in several studies in the Pacific Northwest to organize information from disparate sources at a variety of spatial scales on habitat relationships for wildlife and plant species. Bayesian methods allow the user to calculate the probabilities of outcomes (such as presence of a species) based on the relationships among variables, which may be derived from a combination of empirical data collected via probability and/or non-probability-based sampling, literature reviews, and expert opinion. For example, Marcot (2006) described a Bayesian model for predicting the presence of the Malone jumping slug (*Hemphillia malonei*) based on a combination of effective climate and forest vegetation (as represented by forest ecological zone) at a coarse scale, organic forest floor thickness, woody debris, wet soil patches, and canopy closure of various vegetation layers. Bayesian network models were also used extensively in the Interior Columbia Basin Ecosystem Management Project to examine habitat relationships for several fish and wildlife species (Marcot *et al.* 2001). Lamon & Qian (2008) used Bayesian analysis to combine lake nutrient data collected under a variety of protocols over a time period of 15 years to model a measure of eutrophication (chlorophyll *a*). Their article discusses methods for overcoming two of the common problems in data aggregation: transformation of variables collected with different measurement methods to create common variables across data sets, and gaps in data coverage.

Outputs (estimates) from models are typically associated with some measure of variability. A significant issue in modeling is "propagation of error", or uncertainty about how the variation in the individual model parameters in aggregate affect the variation associated with the model estimates. For example, Ruckelshaus *et al.* (1997) found that variation in dispersal-related mortality of organisms had a disproportionately large effect on overall survival in a spatially explicit population model, a relationship that would not have been revealed without some form of sensitivity analysis. On the other hand, in an exploration of various uncertainty analysis methods for predictive models (using a hydrological water balance model as the subject) Benke *et al.* (2008) found that the variability in predicted annual stream flow was unexpectedly insensitive to variation in the underlying model parameters. Consultation with a statistician for appropriate uncertainty analysis methods is recommended for data aggregation exercises involving models.

Complex models such as the Ecosystem Diagnosis and Treatment (Lestelle *et al.* 2004) model and the All-H Analyzer (Mobrand – Jones & Stokes Associates 2007) model are increasingly used for making regional predictions about salmonids and other populations. Although very powerful, such models have potential shortcomings. First, over-parameterization may result in lack of independence among variables, and may make models relatively insensitive to altered variable values and spatial extents. Second, such models tend to be extremely difficult to validate or verify because of the complexities of assumptions and relationships. Whenever such models are used, it is wise to test them against various other models that are based on alternative conceptual frameworks, assumptions, and datasets.

Meta-analysis

First used in the social and medical sciences, meta-analysis has recently become a common analytical tool in the ecological sciences. Meta-analysis consists of a set of statistical methods designed to draw rigorous inferences from multiple studies (Gurevitch *et al.* 2001). The salient features of meta-analysis are the determination of treatment effects across studies measuring similar variables, the ability to calculate confidence limits around mean treatment effects (in other words, explicit incorporation of sampling variance), and tests for consistency in the trends or sizes of effects among studies (Gurevitch *et al.* 2001). A key underlying assumption in meta-analysis is that combining effect sizes from individual studies (which are considered independent estimates of the true effect size, subject to random variation) provides a better estimate of the true effect than any one study by itself (Gates 2002). Meta-analysis differs from traditional synthesis methods (such as narrative reviews or vote-counting) in that it provides a formalized framework of analytical techniques intended to confer statistical rigor and minimize bias (Gates 2002).

Meta-analysts commonly use one or more of a standard set of measures of effect size (standardized mean difference or Hedge's *d*, the log response ratio, correlation coefficient, or odds ratio; Gurevitch *et al.* 2001), each of which are more or less appropriate to certain types of analyses. For a summary, see Lajeunesse & Forbes (2003). Some authors feel that the use of standardized meta-analysis metrics can lead to erroneous results (i.e., one suite of metrics does not fit all analyses), and recommend construction of metrics specific to the needs of individual analyses (Gurevitch *et al.* 2001). To assess differences in treatment effects, methods analogous to ANOVA and regression have been developed for meta-analysis, as have resampling methods such as bootstrapping (Adams *et al.* 1997).

Meta-analysis is not without its critics. Concerns about the reliability of these techniques (especially in ecological applications) include lack of sufficient studies to cover the spectrum of environments across which effects are being assessed, bias in the selection of studies to include in the meta-analysis, and "research bias" wherein researchers select for study only those features that will likely yield significant results (Lortie & Callaway 2006). In his review of quantitative syntheses using meta-analysis, Gates (2002) makes the following observations about selection of studies for inclusion:

- Meta-analysis of non-randomized data can yield biased results; this tendency should be factored in throughout the analysis and interpretation of results.

- *All* appropriate studies should be initially considered for a meta-analysis, including (or maybe especially) those that do not show significant effects. In addition to reports from peer-reviewed journals, grey literature, foreign-language journals, and unpublished studies should be considered. The rationale is that relying solely on peer-reviewed publications may bias the selection of studies toward those that show significant effects, or that support prevailing hypotheses.
- Techniques for identifying selection bias, such as the construction of funnel plots, should be used.
- Testing for the sensitivity of the meta-analysis to the inclusion criteria (e.g., elimination of unpublished studies) should be conducted.

In addition, Gurevitch *et al.* (2001) caution that non-independence among published studies may occur if different aspects of a study are published separately.

While meta-analysis techniques are commonly used to synthesize published research results (e.g., Englund *et al.* 1999; Paquette *et al.* 2006; Marczak *et al.* 2007) they may also be used to combine data sets. For example Anthony *et al.* (2006) used meta-analysis on nonprobability based data to determine the status and trends of northern spotted owl (*Strix occidentalis caurina*) demography (see also Lint *et al.* 1999; Boyce *et al.* 2005). Ojeda-Martinez *et al.* (2006) used meta-analysis of long-term monitoring data to assess the conservation benefits of marine reserves to protected fish.

Part 3. Data Comparability and Other Issues

Monitoring programs initiated or organized by agencies, natural resource and environmental organizations, universities, volunteers, students, and citizen groups have increased significantly over the past 40 years in the US as well as world-wide. Benefits of such programs include heightened public awareness, increased data availability, enhanced collaboration with professionals, and decreased monitoring costs (Gouveia *et al.* 2001; Goffredo *et al.* 2004). Several authors contend that adequately trained volunteers using properly designed, local monitoring schemes can produce locally relevant results that can be as reliable as those derived from professional monitoring programs (e.g., Fore *et al.* 2001; Danielsen *et al.* 2005). Rock & Launten (1996) suggest that if monitoring and sampling schemes are properly designed to allow for data comparison and validation with standard research methods, the research community and government agencies could have access to datasets not otherwise available.

There are several issues related to the use and integration of data collected by all these groups, including professionals (Sharpe & Conrad 2006), that need to be considered before any data are combined and analyzed. These issues include data credibility and reliability (Evans *et al.* 2001; Canfield *et al.* 2002; Hanson 2006) data inconsistencies over time and among observers (Darwall & Dulvy 1996; Rieman *et al.* 1999), non-comparability of data (Boyce *et al.* 2006), insufficient sample sizes (Gouveia *et al.* 2001), differences in sampling effort (Cao *et al.* 2002; Fayram *et al.* 2005; Smith & Jones 2008), data completeness (e.g., low sampling frequency and short time-frames; Rieman *et al.* 1999; Gouveia *et al.* 2001), and incomplete spatial and/or temporal coverage of data (Goffredo *et al.* 2004; Smith & Michels 2006). Another concern, particularly with citizen groups, is whether community/volunteer based programs can put systems in place to maintain quality control and to quickly identify and fix problems when they occur (Sharpe & Conrad 2006). It is important to note that data collected by professional entities can also reflect problems instituting and maintaining quality control.

Several strategies can be used to help ensure data are of high quality and accuracy (Table 5). The following discussion focuses on a few key points. Studies have also been conducted to determine comparability of protocols and data for natural resource surveys; see Appendix B for a summary.

Table 5. Strategies used by various environmental monitoring organizations to address data credibility, non-comparability of results, and data completeness. Modified from Gouveia *et al.* 2001.

Concern	Strategies or practices
Data Credibility	Develop and adhere to quality assurance and quality control plans Collaborate with statistical and scientific community Create and maintain metadata Use sensors and standard data collection methods Conduct parallel testing; data and procedure audits Provide regular training and quality assurance audits to maintain and improve skills Quickly identify and correct problems in sampling procedures and data collection
Data comparability	Create and maintain metadata Collaborate with statistical and scientific community Create training, and protocol manuals data collection activities Use and calibrate sensors Use standard data collection methods Conduct protocol comparison studies
Data completeness	Involve stakeholders and partner groups through feedback from official entities, scientists, and decision makers Maintain good leadership and continuity of programmatic knowledge over time Maintain adequate technical and financial support for monitoring activities

Data Comparability

Significant issues arise when data to be aggregated are collected through the use of differing methods, site extents, sampling or index periods, or survey objectives. Differences between survey methods were discussed in Part 1. Other considerations are the subject of this section. Before undertaking an aggregation, the analyst must consider factors that may affect the comparability of data, starting with the framework of the objectives under which the data were initially gathered. For example, if the objective of a fish sampling event is to assess the abundance of sport fish, the data are unlikely to be sufficient for non-game fish. On the other hand, if the fish are collected for the purpose of studying fish taxonomy, some species may be ignored and the design may be nonprobability-based. In both cases, making a regional assessment of the entire fish assemblage would be problematic.

Combining monitoring data from different seasons can hinder comparisons among sites. Typically, regional assessments focus on relatively narrow sampling periods, such as summer base flows for fish assemblages (Meador *et al.* 1993; Hughes & Peck 2008) and fall turnover for lakes or spring snow melt for streams when assessing surface water acidification (Baker et al. 1990). Such index periods overlook natural seasonal differences and may distort some regional patterns and clarify others. This is a particularly important issue in salmonid monitoring because significant year-to-year differences in adult returns are typical in anadromous salmonids.

Because most biological species are distributed in a patchy manner across their environments, the extent of the area sampled and the sampling intensity are critical factors to compare when determining whether data from different surveys can be combined (Pont *et al.* 2006). For example, recent studies of electrofishing in US streams have indicated that a distance equal to 40 times the mean wetted width of the stream is needed to detect all but the rarest species 90% of the time (Hughes & Peck 2008; LaVigne *et al*. 2008b). Combining data from studies with differing fish sampling distances would therefore produce varying results simply because of the protocols used (LaVigne *et al.* 2008a). Similar concerns arise when comparing macroinvertebrate samples from different collectors. Li *et al.* (2001) determined that multiple point samples are needed for each site, and Cao *et al.* (2002) found that several hundred individuals must be identified for precise estimates of assemblage similarity or dissimilarity. Therefore, the number of samples, the area sampled, the sample location, and the number of individuals processed can affect the comparability of macroinvertebrate samples for regional assessments (Gerth & Herlihy 2006; Blocksom *et al.* 2008; Stoddard *et al.* 2008).

Likewise, combining data that were collected with different sampling equipment may confound regional assessments (Pont et al. 2006; Hughes & Peck 2008; Stoddard et al. 2008). For example, passive sampling gear, such as traps or gill nets, produce different fish species results than active electrofishing or seining of lakes (Vaux *et al.* 2000) or rivers (LaVigne *et al.* 2008a). Similarly, for stream macroinvertebrates, the mesh size of nets must be comparable among studies. Bonar *et al.* (in press) and Bonar & Hubert (2002) encourage standardization of fish sampling methods to maximize data comparability and sharing in the US, similar to what exists in the European Union (CEN 2003).

The two primary issues in data comparability, whether different protocols produce equivalent values, and whether different observers using the same protocol made equivalent measurements or assessments, have been assessed for a variety of protocols and methods (Table B-1 in Appendix B). Comparability varied among methods. Blocksom *et al.* (2008) found that single and multiple habitat sampling protocols did not produce interchangeable data. Herbst & Silldorf (2006) found similarities between three stream macroinvertebrate bioassessment methods and suggested that the data could be used interchangeably. Fiala *et al.* (2006) compared four methods of forest canopy cover estimation and developed regression equations that could be used to convert data between methods used in similar conifer forests. Observer differences were also documented in comparability studies. Roper *et al.* (2008) found that only one-third of field crews using a stream classification in a controlled study agreed on stream channel types. Likewise, Kauffman *et al.* (1999) reported greater error in subjective estimates of habitat indicators than in quantitative measurements.

Results from studies comparing the quality of data collected by volunteers or students *versus* professionals are generally mixed (e.g., Obrecht *et al.* 1998; Fore *et al.* 2001; Engel & Voshell 2002; Brandon *et al.* 2003; Goffredo *et al.* 2004; Galloway *et al.* 2006; Hanson 2006; see Table B-2 in Appendix B). Results often varied by monitoring task. Nicholson *et al.* (2002) found volunteer data for stream turbidity were statistically different from professionally collected data, while no differences were seen in data for electrical conductivity or pH. In a comparison of volunteers vs. professional botanists in an assessment of tree and shrub species frequency of occurrence, Brandon *et al.* (2003) found that for 12 out of 20 species, volunteers and botanists

recorded presence and abundance similarly. The other 8 species not recorded similarly were known to be difficult for non-professionals to differentiate (Brandon *et al.* 2003). In general, non-professional plant identifications have been found to be more accurate at the genus level than the species level (Bloniarz & Ryan 1996; Brandon *et al.* 2003). The use of subjective measures such as abundance categories has also been shown to vary significantly between volunteers and professionals (Evans *et al.* 2001; Foster-Smith & Evans 2003).

Metadata

Metadata records should be developed and maintained for all datasets and sampling locations (Boyce *et al.* 2006). The Federal Geographic Data Committee (FGDC) provides comprehensive metadata standards and guidelines that are used by federal agencies and other to prepare documentation for spatial (FGDC 1998) and biological data (FGDC Biological Data Working Group and USGS Biological Resources Division 1999). Inclusion of descriptive metadata and documentation of scientific processes (process metadata) used to produce data sets from raw data will also facilitate their use by others and enable users to build data reliability indicators (Gouveia *et al.* 2001; Ellison *et al.* 2006). Rigorous metadata documentation includes a description of the data, the sampling design and data collection protocols, quality control procedures, preliminary processing, derivatives or extrapolations, estimation procedures, professional judgment used, and any known anomalies or oddities of the data (NRC 1995). The National Research Council (NRC 1995, page 4) recommended that metadata should "*explicitly describe all preliminary processing associated with each data set, along with its underlying scientific purpose and its effects on the suitability of the data for various purposes.*" Further the NRC recommended that the metadata should describe and quantify the statistical uncertainty resulting from each processing step (NRC 1995). The NRC (1995) found that ideally, metadata should provide enough detail to allow users unfamiliar with the data to back track to earlier versions of the data so that they can perform their own processing or derivations as needed. Ellison *et al.* (2006) described an analytic web that provides complete and precise definitions of scientific processes used to process raw data and derived data sets. Analytic webs allow validation of datasets by creating an internet-accessible audit trail of the process used to transform raw data into the available form (Ellison *et al.* 2006).

Integrating Datasets

Once objectives for aggregating data have been decided and datasets chosen, several issues should be considered before the datasets are actually combined into a new dataset for analysis. Ensuring data integrity when integrating data into a new, larger dataset can be a considerable task (McLaughlin *et al.* 2001; Astin 2006). Detailed metadata can make merging data sets easier and identify possible incompatibilities. Quality assurance and quality control techniques can be used to assess data credibility (Savan *et al.* 2003). Data validation frameworks with tools for checking data quality are helpful (Gouveia *et al.* 2001). Problems encountered could include, but are not limited to:

- data sets that are not kept electronically in their entirety (e.g., location information or date of collection may be kept on hard copies),

- data formats (e.g., metric vs. English measurements, different decimal places) and file types may be inconsistent or incompatible,
- data fields with the same name may not contain the same type of data or information (e.g., "species" may variously include common names, scientific names, or acronyms), and
- species may not be identified to the same taxonomic level (e.g., species, subspecies, or variety may not be recorded).

All possible errors in the datasets also must be checked, corrected if possible, or deleted. If problems are found, data sets that include both raw numbers and calculated values can help identify the source of errors and to make corrections (Lawless & Rock 1998).

Accepting data at face value can introduce unknown errors into analyses and interpretations. In a project to combine federal and state agency data (including historical data) on Great Lakes stream fishes, McLaughlin *et al.* (2001) admitted that assuming data received from agencies was free from errors was a less than desirable approach, but was necessary due to limited resources. Pont *et al.* (2006) showed an approach for dealing with questions about the quality of historical data in an aggregation involving European stream fish species and abundances for 5,252 sites, 12 nations, and 24 years. They rejected data not collected by a standard electrofishing method, during low flows, over a sufficient area, or in multiple passes at the same site. They also retained variables indicating whether the site was fished by wading or boat, the size of area fished, and whether the site was sampled completely or only nearshore, so that the effects of those variables could be evaluated. The aggregation resulted in a predictive model for assessing fish assemblage condition at all 5,252 sites Pont *et al.* (2006).

Part 4. Summary and Conclusions

The goal of this report has been to balance the presentation of tools for aggregating spatial environmental data from different sources and scales, with cautions about the statistical complexities of doing so. Heightened interest in evaluating the success of policies for managing natural resources and protecting the environment makes it increasingly likely that disparate information will be used in assessing status and trends of ecosystems on a regional basis. The IMST hopes that the techniques presented here, along with the caveats for their use, will help inform such efforts. Furthermore, the IMST hopes that raising the awareness of the difficulties inherent in combining data from different sources will increase the likelihood that future sampling efforts will be planned to accommodate aggregation.

Considerations when Planning to Aggregate Data

A first step in approaching data aggregation is to understand the properties of the data and the underlying sampling design. This will determine the basis from which conclusions or inferences can be drawn and what techniques can be used. Important questions to be answered include:

- *Are the scales at which the studies were conducted and the variables contained in the datasets, actually relevant to the question(s) being answered by the aggregation?* This step involves scrutinizing the available variables to determine if the information and scales they represent can be appropriately applied to the scale and information needed for the aggregation.
- *Do the data come from probability-based or nonprobability-based sampling, or a combination?* Data from probability-based sampling designs are often easier to combine in a "summing up" sense without distorting relationships among the variables.
- *What is the "geography" of the data? Do the studies completely cover the geographic area of interest, or are there gaps (environmental or spatial)? Are the spatial patterns of environments within the area likely to confound the interpretation of results?* Making inferences to unsampled areas that are environmentally different from the sampled area or assuming all variables operate similarly across space can lead to erroneous conclusions.

Another important step in planning an aggregation is to determine the basis from which conclusions are drawn from the data. In *design-based* approaches, data from probability-based sampling designs are used, and the basis for making inferences is the design of the study itself. Because variability can be quantified in design-based approaches, they are frequently used in studies aimed at making estimates (e.g., of population numbers or water quality parameters) with a known uncertainty factor. However, design-based approaches cannot legitimately be used to "predict", that is, to make inferences about areas that were not sampled (e.g., using water quality data from one river basin to predict water quality in a different river basin). A design-based approach to aggregation can only be used when there are common variables among datasets, and the samples can be combined statistically into a single sample. If these conditions do exist, a design-based aggregation can be a powerful tool to make broad-scale inferences.

In model-based approaches, data may be from either probability-based or nonprobability-based sampling designs, and the basis for drawing conclusions is some type of model. Since models lack the ability to rigorously quantify variability, the reliability of their estimates of uncertainty is unknown. However, models have the advantage of being able to incorporate a wider range of sampling designs, and can be more helpful in understanding relationships among variables. In addition, models can be used to make predictions to unsampled areas, as long as there is some reliable environmental information (e.g., GIS representation of vegetation types) available to make comparisons between sampled and unsampled areas. When using models, it is important to keep in mind that model outputs are a function of the assumptions and conceptual frameworks that underlay the model structure, therefore, consideration of alternative assumptions and frameworks is recommended.

Potential Problems in Data Aggregations

The primary issue in data aggregation, whether simply adding results of similar studies together or combining studies with different variables at different scales in a complex model, is that the process of summing and transforming variables can cause the relationships among them to change or become obscured. This necessitates tools for both discerning if and how the relationships are changing and for mitigating problems.

Distorted conclusions from aggregated data can arise from several sources, including the sampling designs themselves, the process of "summing" data in alternative ways, the process of spatially "scaling up" or changing the geographical boundaries represented by the data, and hidden influences in the environment that are not taken into account in sampling. Table 6 summarizes some of the potential problems, which were discussed in detail in Parts 1 and 3.

Methods for Aggregating Data

The ability to aggregate data appropriately depends on the designs of the studies under which data were collected. Because of the complexities, consultation with a statistician with experience in aggregation techniques is an important first step, especially when combining probability-based and nonprobability-based data.

Aggregating data from probability samples is relatively straightforward and basically involves creating a single probability sample from the component studies. In order for probability samples to be combined, they must have commonality among variables of interest and sufficient information about sampling frames and sample site selection methods to allow comparisons to be made. Methods for aggregating probability samples include combining weighted estimates, post-stratification, and direct combination into a single sample (Cox & Piegorsch 1996; Olsen *et al.* 1999)

Table 6. A summary of some of the problems that can be encountered when aggregating data.

Problem encountered	What it is	References for more information
Simpson's Paradox	Relationships between attributes appear to change (or even reverse) depending on how a population and its attributes are stratified. Occurs with discrete data in descriptive statistical analyses.	Wagner 1982; Cohen 1986; Thomas & Parresol 1989
Change of Support Problem.	Occurs when observations are made on one spatial scale but the process of interest is operating at different spatial scale. Can create inference problems.	Gotway & Young 2002, Gotway Crawford & Young 2005
Modifiable Areal Unit Problem	Occurs when changes in the size, configuration, and number of groupings of data alter the apparent relationships. May obscure actual relationships.	Openshaw and Taylor 1979; Openshaw 1983; Jelinski and Wu 1996
Ecological Correlation	Correlations occur between group means as opposed to individual means.	Robinson 1950; Clark & Avery 1976
Ecological Fallacy	Occurs when the relationships between group means is inferred to individuals, leading to false conclusions about individuals	Johnson & Chess 2006
Pseudoreplication	Occurs when the scale of the experimental unit is misidentified and the number of independent replicates appears larger than it really is. Observations are actually interdependent. Variability is misrepresented, and statistical power is overstated.	Hurlbert 1984; Heffner *et al.* 1996; Death 1999; Miller & Anderson 2004
Lurking Variables	Occurs when the presence of an unknown or non-measured variable affects the relationships between measured variables	Cressie 1996
Spatial Auto-correlation	Occurs when variables have a tendency to aggregate spatially. May lead to erroneous conclusions about causes of distribution.	Fortin *et al.* 1989; Lichstein *et al.* 2002; Diniz-Filho *et al.* 2003
Cross-scale Correlation	Occurs when there is correlation between variables at different spatial scales. May lead to erroneous conclusions about which variable/spatial scale is most significant.	Battin & Lawler 2006; Lawler & Edwards 2006
Data Incomparability	Occurs when data are collected through use of differing methods, site-scale designs, indicators, index periods, spatial scales, or survey objectives.	Bonar & Hubert 2002; Cao *et al*. 2002; Gerth & Herlihy 2006; McDonald *et al.* 2007; Hughes & Peck 2008

Combination of probability-based data with nonprobability-based data has significant limitations that must be factored into the analysis. The primary problem is that quantitative estimates of variation and uncertainty cannot be calculated from nonprobability-based data, so the validity of the results cannot be quantified. The nature and objectives of the aggregation will determine how severe a problem this may be. Methods for combining probability and nonprobability data include those that treat the nonprobability data as though it was probability-based (e.g., pseudo-random and stratified calibration approaches; Overton *et al*. 1993), Brus & de Guijter approach (Brus & de Gruijter 2003), models, and meta-analysis.

In any of these methods, ensuring the comparability and quality of data is essential. Issues such as consistency in the use of methods and equipment, timing of sampling, and equivalency in definitions of variables must be addressed. Adequate documentation (metadata) of methods, dataset structure, and sampling locations is also important.

Concluding Observations

Based on IMST's review of the issues related to aggregating data to assess environmental conditions, the IMST makes the following observations:

- The potential for future aggregation should be considered in the design of data collection efforts, whether they are broad scale surveys or small research. This would include rigorous documentation of study objectives, assumptions, sampling design, variable definitions, implementation records, and database structure.
- Further use of the "Master Sample" concept (a standardized spatially balanced probability-based sampling design described in Larsen et *al*. [2008] as a basis for investment in integrated data collection) should be considered by monitoring and research groups.
- The services of a statistician with experience in data aggregation methods should be obtained when planning data aggregation projects. Early consultation is recommended, especially at the stages of setting objectives, evaluating studies for inclusion in the aggregation, and deciding which methods to use.
- In all analyses, uncertainty should be quantified if possible. In the use of methods (such as some models) where it is not possible, alternative conceptual frameworks and sets of assumptions, as well as model validation, should be considered.

Glossary of Terms

Areal centroid – the geographic center of an area.

Basal area growth – growth in the cross sectional area of a tree stem.

Bootstrapping – a computer-intensive (i.e., often requires a large amount of computation) method used to resample data to determine statistical significance.

Change of support problem (COSP) – a concept in geostatistics, used to describe problems associated with making inferences when the observed variable and the actual feature or process of interest are at different scales (e.g., inferring a spatial continuum of temperature from point data). "Support" refers to the geometric size (or volume), shape, and spatial orientation of the area associated with a measurement, which is altered when inferences are made to a different scale.

Cluster sampling – a probability-based sampling technique used for isolated or natural groupings of population units that may not be equally distributed across the landscape or over time.

Continuous spatial domain – a contiguous area or region.

Contouring – the construction of contour lines through points with similar values.

Convenience sampling – see *Opportunistic sampling*

Correlation coefficient – a statistical measure of the linear association between two variables (e.g., Pearson's correlation coefficient).

Cross-scale correlation – occurs when variables measured at different spatial scales are statistically correlated.

Data aggregation – combining data from different studies or surveys.

Diameter class – a standardized grouping of tree stem diameter sizes taken at breast height (1.4 m above the ground). For example, a 14 cm diameter class includes trees ranging from 13.6 – 14.5 cm in diameter at breast height.

Dual-frame estimation – a dual-frame design is used to combine two sampling frames in the same survey to offer coverage rates that may exceed that of any single frame. See *Frame sampling.*

Ecological correlation – a correlation between two variables that are group means, in contrast to a correlation between two variables that describe individuals.

Ecological fallacy –an error in the interpretation of statistical data, whereby inferences about the nature of individuals are based on aggregate statistics collected for a group to which the individuals belong.

Effect size – in meta-analysis, the magnitude of a presumed relationship among variables and conditions or treatments. Various indices are used to measure treatment effects, including Hedge's *d* and the log response ratio.

Element, population – a member of a population, where a statistical population consists of a total set of elements (e.g., organisms, habitats, substrates, streams, etc.) under study.

Frame, Sampling – sampling frame is the population of units or individuals from which samples are taken. Ideally the sample frame accurately represents the target population.

Geostatistics – a subset of statistical analyses used in geology.

Gradient studies – studies in which sampling occurs along a known or suspected environmental gradient to ascertain relationships between the gradient and the response variables of interest.

Hedge's *d* – one of several indices used to measure the magnitude of a treatment effect in meta-analysis. See *Effect size*

Inclusion probability –generally, the probability that a specific sample will be selected from a population. A *first-order inclusion probability* is the likelihood that a single sample will be chosen from the population. A *second-order inclusion probability* is the likelihood that a pair of elements will be chosen from the population.

Incompatible or misaligned zones – occurs when data from different sources do not coincide because of differences in spatial locations or scales.

Invariant –in statistics, the property of a relationship remaining unchanged when a particular transformation is applied to it: a quantity or expression that is constant throughout a certain range of conditions.

Log response ratio – one of several indices used to measure the magnitude of a treatment effect in meta-analysis. See *Effect size*

Lurking (or hidden) variable – a variable that has an important effect on the relationship observed in a study, but is not included among the measured variables.

Kriging – a statistical interpolation technique used to estimate the value of an unsampled point or cell from adjacent sampled points or cells. Block kriging is based on grid cells and point kriging is based on points.

Meta-analysis – a statistical technique used to combine results from multiple independent studies that address the same topic.

Metadata – descriptions and documentation of study sites, plot locations, scientific procedures, experimental designs, and calculations.

Modifiable areal unit problem (MAUP) – from geographical statistics, a problem encountered when the alternate sizes, configuration, or numbers of groupings in the data cause changes in the observed relationships among variables. The MAUP is a type of change of support problem. See *Change of support problem*

Multimetric – several measurable characteristics combined for a biological assemblage

Nonprobability-based (non-random) sampling – sampling without a known, nonzero probability of a particular population element being included in the sample. Inferences about the population with a known level of confidence cannot be drawn directly from the sample, because variance cannot be reliable estimated. See *Convenience sampling, Observational studies,Ppurposive searches, and Gradient studies*

Non-random sampling – see *Nonprobability-based sampling*

Observational studies – a nonprobability-based sampling design in which the observer does not assign subjects to treatment or control groups (e.g., *in situ* habitat studies)

Opportunistic sampling – a nonprobability based sampling technique in which sampling points are non-randomly chosen in an unstructured manner from the frame, often based on accessibility or convenience. Sometimes called *ad hoc, convenience,* or *grab* sampling.

Post-stratification – data analysis technique in which samples are grouped into homogenous strata (see *Stratified random sampling*) after data have been collected. It is possible to increase precision through post-stratification by re-weighting the sample so that the stratum weight in the sample matches the stratum proportion in the population.

Probability-based (random) sampling – sampling in which every member of the frame has a known, nonzero probability of being chosen (the probabilities are not necessarily equal for all sample points). Because variance can be calculated from probability-based samples, population estimates can be made with a known measure of confidence. See *Cluster sampling, Simple random sampling, Stratified random sampling, Systematic sampling, and Spatially balanced sampling.*

Pseudoreplication – a feature of a sampling design in which data are treated as independent observations when they are actually interdependent; observations are not true replicates and statistical power may therefore be overstated.

Purposive searches – a nonprobability-based technique that focuses sampling on sites where the variable of interest is most likely to be found, based on expert knowledge or previous findings.

Radial growth –growth in the radius of a tree bole, usually measured at 1.4 m above the ground).

Random sampling – See *Probability-based sampling*

Regression coefficient – a statistical value indicating the strength and direction of a relationship between predictor and response variables.

Sample Frame – See *Frame, sampling*

Scale – in a general sense, the extent (i.e., relative length, area, or size) and grain or resolution of spatial information. On maps it specifically refers to the ratio between map length and ground distance (e.g., 1:63,360 indicates a scale of 1 map inch to 63,360 ground inches or 1 mile).

Simple random sampling – sample design in which samples are chosen from a frame randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of *k* individuals has the same probability of being chosen for the sample as any other subset of *k* individuals

Simpson's Paradox – a phenomenon in statistical analysis in which the apparent association of variables appears to change or reverse when they are combined in alternative groupings.

Spatial autocorrelation – occurs when measurements of spatial variables are not independent, i.e., values of a variable depend on or can be predicted from values at nearby sites.

Spatially balanced sampling – survey designs that are probability-based and in which the spatial distribution of samples reflects that of the population.

Stratified random sampling – a probability-based sampling technique that groups members of the population into relatively homogeneous subgroups before sampling. The strata should be mutually exclusive: every element in the population must be assigned to only one stratum. Then random or systematic sampling is applied within each stratum. The sample number is usually relative to the size of the stratum.

Synoptic sample – a sample from the entire population, i.e., no portions of the sampling frame were excluded.

Systematic sampling – a probability-based sampling technique that selects equally spaced sampling points on a grid or transect. The first point is the only random point.

Trend surface analysis – a global surface-fitting procedure used in GIS and environmental science in which characteristics of a surface, represented by a regular grid of points, are approximated from unevenly distributed points.

Unequal probability sampling – a sampling technique in which the sampling probabilities are based on knowledge of some type of structure in the population, such as proportionality.

Literature Cited

Adams DC, Gurevitch J, Rosenberg MS (1997) Resampling tests for meta-analysis of ecological data. *Ecology* 78(5):1277–1283.

Allison VJ, Goldberg DE (2002) Species-level versus community-level patterns of mycorrhizal dependence on phosphorous: An example of Simpson's Paradox. *Functional Ecology* 15(3): 346–352.

Amrhein CG (1995) Searching for the elusive aggregation effect: evidence from statistical simulations. *Environment and Planning A* 27(1): 105–119.

Anthony RG , Forsman ED, Franklin AB, Anderson DR, Burnham KP, White GC, Schwarz CJ, Nichols JD, Hines JE, Olson GS, Ackers SH, Andrews LS, Biswell BL, Carlson PC, Diller LV, Dugger KM, Fehring KE, Fleming TL, Gerhardt RP, Gremel SA, Gutiérrez RJ, Happe PJ, Herter DR, Higley JM, Horn RB, Irwin LL, Loschl PJ, Reid JA, Sovern SG (2006) Status and trends in demography of northern spotted owls 1985–2003. *Wildlife Monographs* 163(1): 1–48.

Astin LE (2006) Data synthesis and bioindicator development for nontidal streams in the interstate Potomac River basin, USA. *Ecological Indicators* 6(4): 664–685.

Baker LA, Kaufmann PR, Herlihy AT, Eilers JM (1990) *Current Acid Base Status of Surface Waters.* State of Science Report # 9, National Acid Precipitation Assessment Program, Washington DC.

Battin J, Lawler J (2006) Cross-scale correlation and the design and analysis of avian habitat selection studies. *Condor* 108(1): 59–70.

Benda L, Miller D, Andras K, Bigelow P, Reeves G, Michael D (2007) NetMap: A new tool in support of watershed science and resource management. *Forest Science* 53(2): 206–219

Benke KK, Lowell KE, Hamilton AJ (2008) Parameter uncertainty, sensitivity analysis and prediction error in a water-balance hydrological model. *Mathematical and Computer Modelling* 47(11–12): 1134–1149.

Blocksom KA, Autrey BC, Passmore M, Reynolds L (2008) A comparison of single and multiple habitat protocols for collecting macroinvertebrates in wadeable streams. J*ournal of the American Water Resources Association* 44(3): 577–593.

Bloniarz DV, Ryan HDP III (1996) The use of volunteer initiatives in conducting urban forest resource inventories. *Journal of Arboriculture* 22(2): 75–82.

Bonar S, Hubert WA (2002) Standard sampling of inland fish: Benefits, challenges, and a call for action. *Fisheries* 27(3): 10–16.

Bonar S, Hubert WA, Willis D, eds. (in press) *Standard Methods for Sampling North American Freshwater Fishes*. American Fisheries Society, Bethesda, Md.

Boudreau SA, Yan ND (2004) Auditing the accuracy of a volunteer-based surveillance program for an aquatic invader *Bythotrephes. Environmental Monitoring and Assessment* 91(1–3): 17–26.

Boyce MS, Irwin LL, Barker R (2005) Demographic meta-analysis: synthesizing vital rates for spotted owls. *Journal of Applied Ecology* 42(1): 38–49.

Boyce D, Judson B, Hall S (2006) Data sharing – A case of shared databases and community use of on-line GIS support systems. *Environmental Monitoring and Assessment* 113(1–3): 385–394.

Brandon A, Spyreas G, Molano-Flores B, Carroll C, Ellis J (2003) Can volunteers provide reliable data for forest vegetation surveys? *Natural Areas Journal* 23(3): 254–261.

Bray GS, Schramm HL Jr (2001) Evaluation of a statewide volunteer angler diary program for us as a fishery assessment tool. *North American Journal of Fisheries Management* 21(3): 606–615

Brenden TO, Clark RD, Cooper AR, Seelbach PW, Wang L (2006) A GIS framework for collecting, managing, and analyzing multiscale landscape variables across large regions for river conservation and management. Pages 49–74 in: *Landscape influences on stream habitats and biological assemblages* (eds Hughes RM, Wang L, Seelbach PW) Symposium 48, American Fisheries Society, Bethesda, Md.

Bromenshenk JJ, Preston EM (1986) Public participation in environmental monitoring: A means of attaining network capability. *Environmental Monitoring and Assessment* 6(1): 35–47.

Brus DJ, de Gruijter JJ (2003) A method to combine non-probability sample data with probability sample data in estimating spatial means of environmental variables. *Environmental Monitoring and Assessment* 83(3): 303–317.

Canfield DE Jr, Brown DC, Bachmann RW, Hoyer MV (2002) Volunteer lake monitoring: Testing the reliability of data collected by the Florida LAKEWATCH program. *Lake and Reservoir Management* 18(1): 1–9.

Cao Y, Larsen DP, Hughes RM, Angermeier PM, Patton TM (2002) Sampling effort affects multivariate comparisons of stream assemblages. *Journal of the North American Benthological Society* 21(4): 701–714.

CEN (Comité Européen Normalisation, European Committee for Standardization) (2003) *Water Quality–Sampling of Fish with Electricity*. ICS 13.060.70,65.150. Brussels, Belgium.

Clark WAV, Avery KL (1976) The effects of data aggregation in statistical analysis. *Geographical Analysis* 8: 428–438.

Cohen JE (1986) An uncertainty principle in demography and the unisex issue. *The American Statistician* 40(1): 32–39.

Cox LH, Piegorsch WW (1994) *Combining Environmental Information: Environmetric research in ecological monitoring, epidemiology, toxicology, and environmental data reporting*. Technical Report Number 12. National Institute of Statistical Sciences, Research Triangle Park, NC (accessed April 8, 2008 http://www niss org/technicalreports/tr12 pdf).

Cox LH, Piegorsch WW (1996) Combining environmental information I: Environmental monitoring, measurement and assessment. *Environmetrics* 7(3): 299–308.

Courbois J-Y, Katz SL, Isaak DJ, Steel EA, Thurow RF, Wargo Rub AM, Olsen T, Jordan CE (2008) Evaluating probability sampling strategies for estimating redd counts: An example with Chinook salmon (*Oncorhynchus tshawytscha*). *Canadian Journal of Fisheries and Aquatic Science* 65(9): 1814–1830.

Cressie N (1996) Change of support and the modifiable areal unit problem. *Geographical Systems* 3: 159–180. (Overseas Publishers Association Venezuela)

Daly HE (1973) The steady-state economy: Toward a political economy of biophysical equilibrium and moral growth. Pages 149–174 in: *Toward a Steady-State Economy* (ed Daly HE) Freeman, San Francisco, CA.

Danielsen F, Burgess ND, Balmford A (2005) Monitoring matters: Examining the potential of locally based approaches. *Biodiversity and Conservation* 14(11): 2507–2542.

Darwall WRT, Dulvy NK (1996) An evaluation of the suitability of non-specialist volunteer researchers for coral reef fish surveys Mafia Island, Tanzania — A case study. *Biological Conservation* 78(3): 223–231.

Davies SP, Jackson SK (2006) The biological condition gradient: A descriptive model for interpreting change in aquatic ecosystems. *Ecological Applications* 16(4): 1251–1266.

Death RG (1999) Macroinvertebrate monitoring: statistical analysis. Page 105–117 in: *The use of Macroinvertebrates in Water Management* (ed Winterbourn MJ) Ministry for the Environment, Wellington, New Zealand

Diamond J (2005) *Collapse: How Societies Choose to Fail or Succeed*. Penguin, New York, NY.

Diniz-Filho JAF, Bini LM, Hawkins BA (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography* 12(1): 53-64.

Dorr BS, Burger LW, Barras SC (2008) Evaluation of aerial cluster sampling of double-crested cormorants on aquaculture ponds in Mississippi. *Journal of Wildlife Management* 72(7): 1634–1640.

Edwards TC Jr, Cutler DR, Geiser L, Alegria J, McKenzie D (2004) Assessing rarity of species with low detectability: Lichens in the Pacific Northwest. *Ecological Applications* 14(2): 414–424.

Ellison AM (1996) An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* 6(4): 1036–1046.

Ellison AM, Osterweil LJ, Clarke L, Hadley JL, Wise A, Boose E, Foster DR, Hanson A, Jensen D, Kuseja P, Riseman E, Schultz H (2006) Analytic webs support the synthesis of ecological data sets. *Ecology* 87(6): 1345–1358.

Engel SR, Voshell JR Jr (2002) Volunteer biological monitoring: Can it accurately assess the ecological condition of streams? *American Entomologist* 48(3): 164–177.

Englund G, Sarnelle O, Cooper SD (1999) The importance of data-selection criteria: Meta-Analysis of stream predation experiments. *Ecology* 80(4): 1132–1141.

Evans SM, Birchenough AC, Fletcher H (2000) The value and validity of community-based research: TBT contamination of the North Sea. *Marine Pollution Bulletin* 40 (3): 220–225.

Evans SM, Foster-Smith J, Welch R (2001) Volunteers assess marine biodiversity. *Biologist* 48(4): 168–172.

Fayram AH, Miller MA, Colby AC (2005) Effects of stream order and ecoregion on variability in coldwater fish index of biotic integrity scores within streams in Wisconsin. *Journal of Freshwater Ecology* 20(1): 17–25.

FGDC (Federal Geographic Data Committee) (1998) *Content Standard for Digital Geospatial Metadata* (revised June 1998) Federal Geographic Data Committee, Washington, DC.

FGDC (Federal Geographic Data Committee) Biological Data Working Group and USGS (US Geological Survey) Biological Resources Division (1999) *Content Standard for Digital Geospatial Metadata Part 1: Biological Data Profile*. FGDC-STD-001 1-1999. Federal Geographic Data Committee Washington, DC.

Fiala ACS, Garman SL, Gray AN (2006) Comparison of five canopy cover estimation techniques in the western Oregon Cascades. *Forest Ecology and Management* 232(1–3): 188–197.

Fore LS, Paulsen K, O'Laughlin K (2001) Assessing the performance of volunteers in monitoring streams. *Freshwater Biology* 46(1): 109–123.

Fortin M-J, Drapeau P, Legendre P (1989) Spatial autocorrelation and sampling design in plant ecology. *Vegetatio* 83(1–2): 209-222.

Foster-Smith J, Evans SM (2003) The value of marine ecological data collected by volunteers. *Biological Conservation* 113(2): 199–213.

Galloway AWE, Tudor MT, Vander Haegen WM (2006) The reliability of citizen science: A case study of Oregon white oak stand surveys. *Wildlife Society Bulletin* 34(5): 1425–1429.

Gates S (2002) Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology* 71(4):547-557.

Gelfand AE, Zhu L, Carlin BP (2001) On the change of support problem for spatio-temporal data. *Biostatistics* 2(1): 31–45.

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

Gerth WJ, Herlihy AT (2006) Effect of sampling different habitat types in regional macroinvertebrate bioassessment surveys. *Journal of the North American Benthological Society* 25(2): 501-512.

Goffredo S, Piccinetti C, Zaccanti F (2004) Volunteers in marine conservation monitoring: A study of the distribution of seahorses carried out in collaboration with recreational scuba dives. *Conservation Biology* 18(6): 1492–1503.

Gotway CA, Young LJ (2002) Combining incompatible spatial data. *Journal of the American Statistical Association* 97(458): 632–648.

Gotway Crawford CA, Young LJ (2005) Change of support: an inter-disciplinary challenge. Pages 1–13 in: *Geostatistics for Environmental Applications* (eds Renard P, Demougeot-Renard H, Froidevaux R) Springer-Verlag, Berlin.

Gouveia C, Fonesca A, Câmara A, Ferreira F (2004) Promoting the use of environmental data collected by concerned citizens through information and communication technologies. *Journal of Environmental Management* 71(2): 135–154.

Greenspan A (2008) We will never have a perfect model for risk. *Financial Time*s. 16 March.

Gregoire T (1998) Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research* 28(10):1429–1447.

Griffith GE, Omernik JM, Woods AJ (1999) Ecoregions, watersheds, basins, and HUCs: How state and federal agencies frame water quality. *Journal of Soil and Water Conservation* 54(4): 666–677.

Gurevitch J, Curtis PS, Jones MH (2001) Meta-analysis in ecology. *Advances in Ecological Research* 32: 199–247.

Hale, S.S.; Miglarese, A.H.; Bradley, M.P.; Belton, T.J.; Cooper, L.D.; Frame, M.T.; Friel, C.A.; Harwell, L.M.; King, R.E.; Michener, W.K.; Nicolson, D.T.; Peterjohn, B.G. 2003. Managing troubled data: Coastal data partnerships smooth data integration. *Environmental Monitoring and Assessment* 81(1–3): 133–148.

Halstead LE, Howery LD, Ruyle GB (2000) Comparison of 3 techniques for monitoring use of western wheatgrass. *Journal of Range Management* 53(5): 499–505.

Hansen MH, Madow WG, Tepping BJ (1983) An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association* 78(384): 776–793.

Hanson S (2006) Volunteer vs. agency comparison: *E. coli* monitoring. *The Volunteer Monitor* 18(1): 7 & 12.

Heffner RA, Butler MJ, Keelan RC (1996) Pseudoreplication revisited. *Ecology* 77(8): 2558–2562.

Hendry DF, Ericsson NR (2001) *Understanding economic forecasts*. MIT Press, Boston, Mass.

Herbst DB, Silldorff EL (2006) Comparison of the performance of different bioassessment methods: Similar evaluations of biotic integrity from separate programs and procedures. *Journal of the North American Benthological Society* 25(2): 513–530.

Herger LG, Weiss A, Augustine S, Hayslip G (2003) *Modeling fish distributions in the Pacific Northwest Coast Range Ecoregion using EMAP data.* EPA/910/R-03/000, US EPA Region 10, Seattle, WA.

Herlihy AT, Larsen DP, Paulsen SG, Urquhart NS, Rosenbaum BJ (2000) Designing a spatially balanced, randomized site selection process for regional stream surveys: The EMAP mid-Atlantic pilot study. *Environmental Monitoring and Assessment* 63(1): 95–113.

Hollenhorst TP, Brown TN, Johnson LB, Ciborowski JJH, Host GE (2007) Methods for generating multi-scale watershed delineations for indicator development in Great Lakes coastal ecosystems. *Journal of Great Lakes Research* 33(Sp. Issue 3): 13–26.

Hughes RM, Peck DV (2008) Acquiring data for large aquatic resource surveys: The art of compromise among science, logistics, and reality. *Journal of the North American Benthological Society* 27(4): 837–859.

Hughes RM, Paulsen SG, Stoddard JL (2000) EMAP-Surface Waters: A national, multiassemblage, probability survey of ecological integrity. *Hydrobiologia* 422/423: 429–443.

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54(2): 187–211

IMST (Independent Multidisciplinary Science Team) (2006) *Watershed and Aquatic Habitat Effectiveness Monitoring: A Synthesis of the Technical Workshop* Technical Report 2006-1, Oregon Watershed Enhancement Board, Salem, OR.

IMST (Independent Multidisciplinary Science Team) (2007) *Considerations for the Use of Ecological Indicators in Restoration Effectiveness Evaluation* Technical Report 2007-1, Oregon Watershed Enhancement Board, Salem, Oregon.

Jacobs SE, Cooney CX (1995) Improvement of methods used to estimate the spawning escapement of Oregon coastal natural coho salmon. Annual Progress Report 1993–1994, Fish Research Project, Department of Fish and Wildlife, Portland, OR.

Jelinski DE, Wu J (1996) The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology* 11(3): 129–140.

Johnson BB, Chess C (2006) Evaluating public responses to environmental trend indicators. *Science Communication* 28(1): 64–92.

Johnson SE, Mudrak EL, Beever EA, Sanders S, Waller DM (2008) Comparing power among three sampling methods for monitoring forest vegetation. *Canadian Journal of Forest Research* 38(1): 143–154.

Jones KB, Neale AC, Nash MS, Riitters KH, Wickham JD, O'Neill RV, Van Remortel RD (2000) Landscape correlates of breeding bird richness across the United States mid-Atlantic region. *Environmental Monitoring and Assessment* 63(1): 159–74.

Jones KB, Neale AC, Wade TG, Wickham JD, Cross CL, Edmonds CM, Loveland TR, Nash MS, Riitters KH, Smith ER (2001a) The consequences of landscape change on ecological resources: An assessment of the United States mid-Atlantic region, 1973–1993. *Ecosystem Health* 7(4):229–42.

Jones, KB, Neale AC, Nash MS, Van Remortel RD, Wickham JD, Riitters KH, O'Neill RV (2001b) Predicting nutrient and sediment loadings to streams from landscape metrics: A multiple watershed study from the United States Mid-Atlantic Region. *Landscape Ecology* 16(4): 301–312.

Jones KB, Neal AC, Wade TM, Cross CL, Wickham JD, Nash MS, Edmonds CM, Riitters KH, Smith ER, Van Remortel RD (2005) Multiple scale relationships of landscape characteristics and nitrogen concentrations in streams across a large geographic area. Pages 205–224 in: *Scaling and Uncertainty Analysis in Ecology: Methods and Applications* (eds Wu J, Jones KB, Li H, Loucks OL), Springer, Netherlands.

Karr JR, Fausch KD, Angermeier PL, Yant PR, Schlosser IJ (1986) *Assessing Biological Integrity in Running Waters: A Method and its Rationale*. Special Publication No. 5, Illinois Natural History Survey, Champaign, IL.

Kaufmann PR, Levine P, Robinson EG, Seeliger C, Peck DV (1999) Quantifying physical habitat in wadeable streams. EPA 620/R-99/003, Office of Research and Development, US Environmental Protection Agency, Washington, DC.

Kelley CE, Krueger WC (2005) Canopy cover and shade determinations in riparian zones. *Journal of the American Water Resources Association* 41(1): 37–46.

Kuhn TS (1970) *The Structure of Scientific Revolutions* (2nd ed). University of Chicago Press, Chicago, IL.

Lajeunesse MJ, Forbes MR (2003) Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques. *Ecology Letters* 6(5):448–454.

Lamon EC III, Qian SS (2008) Regional scale stressor-response models in aquatic ecosystems. *Journal of the American Water Resources Association* 44(3): 771–781.

Larsen DP, Olsen AR, Lanigan SH, Moyer C, Jones KK, Kincaid TM (2007) Sound survey designs can facilitate integrating stream monitoring data across multiple programs. *Journal of the American Water Resources Association* 43(2): 384–397.

Larsen DP, Olsen AR, Stevens DL Jr (2008) Using a master sample to integrate stream monitoring programs. *Journal of Agricultural, Biological, and Environmental Statistics* 13(3): 243–254.

LaVigne HR, Hughes RM, Herlihy (2008a) Bioassessments to detect changes in Pacific Northwest river fish assemblages: A Malheur River case study. *Northwest Science* 82(4): 251–258.

LaVigne HR, Hughes RM, Wildman RC, Gregory SV, Herlihy AT (2008b) Summer distribution and diversity of non-native fishes in the main-stem Willamette River, Oregon, 1944–2006. *Northwest Science* 82(2): 83–93.

Lawler J, Edwards TC Jr (2006) A variance-decomposition approach to investigating multiscale habitat associations. *Condor* 108(1): 47–58.

Lawless JG, Rock BN (1998) Student scientist partnerships and data quality. *Journal of Science Education and Training* 7(1): 5–13.

Lestelle LC, Mobrand LE, McConnaha WE (2004) *Information structure of Ecosystem Diagnosis and Treatment (EDT) and habitat rating rules for Chinook salmon, coho salmon, and steelhead trout*. Mobrand Biometrics, Inc. Vashon Island, WA.

Levy PS, Lemeshow S (1999) *Sampling of Populations: Methods and Applications, 3rd edition*. John Wiley & Sons, Inc, New York.

Li J, Herlihy AT, Gerth W, Kaufmann PR, Gregory SV, Urquhart S, Larsen DP (2001) Variability in stream macroinvertebrates at multiple spatial scales. *Freshwater Biology* 46(1):87–97.

Lichstein JW, Simons TR, Shriner SA, Franzreb KE (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs* 72(3): 445–463.

Lint J, Noon B, Anthony R, Forsman E, Raphael M, Collopy M, Starkey E (1999) *Northern Spotted Owl Effectiveness Monitoring Plan for the Northwest Forest Plan.* General Technical Report PNW-GTR-440, USDA Forest Service, Pacific Northwest Research Station, Portland, OR.

Lortie CJ, Callaway, RM (2006) Re-analysis of meta-analysis: Support for the stress-gradient hypothesis. *Journal of Ecology* 94(1):7–16.

Marcot BG (2006) Habitat modeling for biodiversity conservation. *Northwestern Naturalist* 87(1):56–65.

Marcot BG, Holthausen RS, Raphael MG, Rowland MM, Wisdom MJ (2001) Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Management* 153(1–3):29–42.

Marczak LB, Thompson RM, Richardson JS (2007) Meta-analysis: Trophic level, habitat, and productivity shape the food web effects of resource subsidies. *Ecology* 88(1): 140–148.

Mattson MD, Walk MF, Kerr PA, Slepski AM, Zajicek OT, Godfrey PJ (1994) Quality assurance testing for a large scale volunteer monitoring program: The acid rain monitoring project. *Lake and Reservoir Management* 9(1): 10–13

Max TA, Schreuder HT, Hazard JW, Oswald DD, Teply J, Alegria J (1996) *The Pacific Northwest Region Vegetation and Inventory Monitoring System.* Research Paper PNW-RP-493, USDA Forest Service, Pacific Northwest Research Station, Portland, OR.

McCormick FH, Peck DV, Larsen DP (2000) Comparison of geographic classification schemes for Mid-Atlantic stream fish assemblages. *Journal of the North American Benthological Society* 19(3): 385–404.

McDonald LL, Bilby R, Bisson PA, Coutant CC, Epifanio JM, Goodman D, Hanna S, Huntly N, Merrill E, Riddell B, Liss W, Loudenslager EJ, Philipp DP, Smoker W, Whitney RR, Williams RN (2007) Research, monitoring, and evaluation of fish and wildlife restoration projects in the Columbia River Basin: Lessons learned and suggestions for large-scale monitoring programs. *Fisheries* 32(12):582–590.

McGarvey DJ, Hughes RM (2008) Longitudinal zonation of Pacific Northwest (USA) fish assemblages and the species-discharge relationship. *Copeia* 2008(2): 311–321.

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

McLaughlin RT, Carl L, Middel T, Ross M, Noakes DL, Hayes DB, Baylis JR (2001) Potentials and pitfalls of integrating data from diverse sources: Lessons from a historical database for the Great Lakes stream fishes. *Fisheries* 26(7): 14–23.

Meador MR, Cuffney TF, Gurtz ME (1993) *Methods for Sampling Fish Communities as Part of the National Water-Quality Assessment Program*. Open File Report 93-104, US Geological Survey, Raleigh, NC.

Millar RB, Anderson MJ (2004) Remedies for pseudoreplication. *Fisheries Research* 70(2–3): 397–407.

Mobrand – Jones & Stokes (2007) All-H Analyzer (AHA) User Guide – Draft. Updated November 2007. Version 7.3.

Molina R, McKenzie D, Lesher R, Ford J, Alegria J, Cutler R (2003) *Strategic survey framework for the Northwest Forest Plan Survey and Manage Program.* General Technical Report PNW-GTR-573, USDA Forest Service, PNW Research Station, Portland, OR.

Nash RF (1989) *The Rights of Nature: A History of Environmental Ethics*. University of Wisconsin Press, Madison, WI.

NED (Northwest Environmental Data Network) (2005) *Final White Papers and Recommendations from Beyond Ad-Hoc: Organizing, Administrating, and funding a Northwest Environmental Data Network.* Northwest Power and Conservation Council, Portland, OR.

Nerbonne JF, Vondracek B (2003) Volunteer macroinvertebrate monitoring: Assessing training needs through examining error and bias in untrained volunteers. *Journal of the North American Benthological Society* 22(1): 152–163.

Nicholas JW (1997) *The Oregon Plan for Salmon and Watersheds: Oregon Coastal Salmon Restoration Initiative.* State of Oregon, Salem, OR.

Nicholson E, Ryan J, Hodgkins D (2002) Community data – where does the value lie? Assessing confidence limits of community collected water quality data. *Water Science and Technology* 45(11): 193–200

Noon BR, Ishwar NM, Vasudevan K (2006) Efficiency of adaptive cluster and random sampling in detecting terrestrial herpetofauna in a tropical rain forest. *Wildlife Society Bulletin* 34(1): 59–68.

NRC (National Research Council) (1995) *Finding the Forest in the Trees: The challenge of combining diverse environmental data.* National Academy Press, Washington, DC.

O'Connell TJ, Jackson LE, Brooks RP (2000) Bird guilds as indicators of ecological condition in the central Appalachians. *Ecological Applications* 10(6):1706–21.

Obrecht DV, Milanick M, Perkins BD, Ready D, Jones JR (1998) Evaluation of data generated from lake samples collected by volunteers. *Journal of Lake and Reservoir Management* 14(1): 21–27.

Ojeda MM, Sahai H (2002) Design-based sample and probability law-assumed sample: their role in scientific investigation. *International Journal of Mathematical Education in Science and Technology* 33(6): 819–828.

Ojeda-Martinez C, Bayle-Sempere JT, Sanchez-Jerez P, Forcada A , Valle C (2007) Detecting conservation benefits in spatially protected fish populations with meta-analysis of long-term monitoring data. *Marine Biology* 151(3): 1153–1161.

Olsen AR, Peck DV (2008) Monitoring design and extent estimates for the Wadeable Stream Assessment. *Journal of the North American Benthological Society* 27(4): 822–836.

Olsen AR , Sedransk J, Edwards D, Gotway CA, Liggett W, Rathbun S, Reckhow KH, Young LJ (1999) Statistical issues for monitoring ecological and natural resources in the United States. *Environmental Monitoring and Assessment* 54(1): 1–45.

Omernik JM (2003) The misuse of hydrologic unit maps for extrapolation, reporting, and ecosystem management. *Journal of the American Water Resources Association* 39(3): 563–573.

Omernik JM, Bailey RG (1997) Distinguishing between watersheds and ecoregions. *Journal of the American Water Resources Association* 33(5): 935–949.

Omernik JM, Griffith GE (1991) Ecological regions versus hydrologic units: frameworks for managing water quality. *Journal of Soil and Water Conservation* 46(5): 334–340.

Openshaw S (1983) *The Modifiable Areal Unit Problem: Concepts and Techniques in Modern Geography* No. 38, GeoBooks, Norwich.

Openshaw S, Taylor PJ (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Pages 127–144 in: *Statistical Applications in the Spatial Sciences* (ed. Wrigley N), Pion Limited, London, UK.

Overton J, Young T, Overton WS (1993) Using 'found' data to augment a probability sample: procedure and case study. *Environmental Monitoring and Assessment* 26(1):65–83.

Paine RT, Tegner MJ, Johnson EA (1998) Compounded perturbations yield ecological surprises. *Ecosystems* 1(6): 535-545.

Paquette A, Bouchard A, Cogliastro A (2006) Survival and growth of under-planted trees: A meta-analysis across four biomes. *Ecological Applications* 16(4): 1575–1589.

Paulsen SG, Hughes RM, Larsen DP (1998) Critical elements in describing and understanding our Nation's aquatic resources. *Journal of the American Water Resources Association* 34(5): 995–1005.

Paulsen SG, Mayio A, Peck DV, Stoddard JL, Tarquinio E, Holdsworth SM, Van Sickle J, Yuann LL, Hawkins CP, Herlihy AT, Kaufmann PR, Barbour MT, Larsen DP, Olsen AR (2008) Condition of stream ecosystems in the US: An overview of the first national assessment. *Journal of the American Benthological Society* 27(4): 812–821.

Perkins J (2004) *Confessions of an Economic Hit Man*. Penguin Books, New York, NY.

Perry RW, Thill RC (1999) Estimating mast production: an evaluation of visual surveys and comparison with seed traps using white oaks Southern. *Journal of Applied Forestry* 23(3): 164–169.

Philippi T (2005) Adaptive cluster sampling for estimation of abundances within local populations of low-abundance plants. *Ecology* 86(5): 1091–1100.

Piegorsch WW, Cox LH (1996) Combining environmental information. II: Environmental epidemiology and toxicology. *Environmetrics* 7(3): 399–324.

Plafkin JL, Barbour M, Porter K, Gross S, Hughes R (1989) *Rapid Bioassessment Protocols for Use in Streams and Rivers: Benthic Macroinvertebrates and Fish*. EPA 440-4-89-001, US Environmental Protection Agency, Office of Water Regulations and Standards, Washington, DC.

PNAMP (Pacific Northwest Aquatic Monitoring Partnership) (2008) *Pacific Northwest Aquatic Monitoring Partnership*. [Accessed on-line November 25, 2008 http://www.pnamp.org/web/Content.cfm?SectionID=8].

Pont D, Hugueny B, Beier U, Goffaux D, Melcher A, Noble R, Rogers C, Roset N, Schmutz S (2006) Assessing river biotic condition at the continental scale: A European approach using functional metrics and fish assemblages. *Journal of Applied Ecology* 43(1): 70–80.

Ravines RR, Schmidt AM, Migon HS, Rennó CD (2008) A joint model for rainfall-runoff: The case of the Rio Grande Basin. *Journal of Hydrology* 353(1–2): 189–200.

Reeves GH, Hohler DB, Larsen DP, Busch DE, Kratz K, Reynolds K, Stein KF, Atzet T, Hays P, Tehan M (2004) *Effectiveness Monitoring for the Aquatic and Riparian Component of the Northwest Forest Plan: Conceptual Framework and Options.* General Technical Report PNW-GTR-577, USDA Forest Service, PNW Research Stations, Portland, OR.

Rieman BE, Dunham JD, Peterson JT (1999) *Development of a database to support a multiscale analysis of the distribution of westslope cutthroat trout*. Final report to the US Geological Survey, Agreement 1445-HQ-PG-01026BRD, Reston, VA.

Ringold PL, Van Sickle J, Raser K, Schacher J (2003) Use of hemispherical imagery for estimating stream solar exposure. *Journal of the American Water Resources Association* 39(6): 1373–1384.

Robinson A (1950) Ecological correlations and the behavior of individuals. *American Sociological Review* 15(3):351–357.

Rock BN, Lauten GN (1996) K-12th Grade students as active contributors to research investigations. *Journal of Science Education and Technology* 5(4): 255–266.

Rodgers JD (2000) *Abundance of juvenile coho salmon in Oregon coastal streams, 1998 and 1999 Monitoring Program.* Report Number OPSW-ODFW-2000-1, Oregon Department of Fish and Wildlife, Portland, OR.

Roni P, Liermann MC, Jordan C, Steel EA (2005) Steps for designing a monitoring and evaluation program for aquatic restoration. pp 13–34 In: *Monitoring Stream and Watershed Restoration* (ed. Roni P) American Fisheries Society, Bethesda, Md.

Roper BB, Buffington JM, Archer E, Moyer C, Ward M (2008) The role of observer variation in determining Rosgen stream types in northeastern Oregon mountain streams. *Journal of the American Water Resources Association* 44(2): 417–427.

Ruckelshaus M, Hartway C, Kareiva P (1997) Assessing the data requirements of spatially explicit dispersal models. *Conservation Biology* 11(6): 1298–1306.

Savan B, Morgan AJ, Gore C (2003) Volunteer environmental monitoring and the role of the universities: the case of the Citizens' Environment Watch. *Environmental Management* 31(5): 561–568.

Schmitt EF, Sullivan KM (1996) Analysis of a volunteer method for collecting fish presence and abundance data in the Florida Keys. *Bulletin of Marine Science* 49(2): 404–416.

Schooley RL (1994) Annual variation in habitat selection: Patterns concealed by pooled data. *Journal of Wildlife Management* 58(2): 367–374.

Schreuder HT, Williams MS (1995) Design-based estimation of forest volume within a model-based sample selection framework. *Canadian Journal of Forest Research* 25(1): 121–127.

Schreuder HT, Czaplewski R, Bailey RG (1999) Combining mapped and statistical data in forest ecological inventory and monitoring – supplementing an existing system. *Environmental Monitoring and Assessment* 56(3): 269–291.

Schreuder HT, Gregoire TG, Weyer JP (2001) For what applications can probability and non-probability sampling be used? *Environmental Monitoring and Assessment* 66(3): 281–291.

Schindler DE, Augerot X, Fleishman E, Mantua N, Riddell B, Ruckelshaus M, Seeb J, Webster M (2008) Climate change, ecosystem impacts, and management for Pacific salmon. *Fisheries* 33(10): 502–506.

Shapiro MH, Holdsworth S, Paulsen SG (2008) The need to assess the condition of aquatic resources in the US. *Journal of the North American Benthological Society* 27(4): 808–811.

Sharpe A, Conrad C (2006) Community based ecological monitoring in Nova Scotia: Challenges and opportunities. *Environmental Monitoring and Assessment* 113(1–3): 395–409.

Smith TMF (1983) On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society, Series A (General)*, 146(4): 394–403.

Smith KL, Jones ML (2008) Allocation of sampling effort to optimize efficiency of watershed-level ichthyofaunal inventories. *Transactions of the American Fisheries Society* 137(5): 1500–1506.

Smith DR, Michels SF (2006) Seeing the elephant: Importance of spatial and temporal coverage in a large-scale volunteer-based program to monitor horseshoe crabs. *Fisheries* 31(10): 485–491.

Stehman SV, Overton WS (1996) *Spatial Sampling.* Pages 31–63 in: Practical Handbook of Spatial Statistics (ed. Arlinghaus SL), CRC Press, Inc Boca Raton, FL.

Stevens DL Jr (2002) *Sampling Design and Statistical Analysis Methods for the Integrated Biological and Physical Monitoring of Oregon Streams*. OPSW-ODFW-2002-07. The Oregon Plan for Salmon and Watersheds, Oregon Department of Fish and Wildlife, Portland, OR.

Stevens DL Jr, Olsen AR (1999) Spatially restricted surveys over time for aquatic resources. *Journal of Agriculture, Biological, and Environmental Statistics* 4(4): 415–428.

Stevens DL Jr, Olsen AR (2004) Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* 99(465): 262–278.

Stoddard JL, Herlihy AT, Peck DV, Hughes RM, Whittier TR, Tarquinio E (2008) A process for creating multimetric indices for large-scale aquatic surveys. *Journal of the North American Benthological Society* 27(4): 878–891.

Stoddard JL , Peck DV, Paulsen SG, Van Sickle J, Hawkins CP, Herlihy AT, Hughes RM, Kaufmann PR, Larsen DP, Lomnicky G, Olsen AR, Peterson SA, Ringold PL, Whittier TR (2005) *An Ecological Assessment of Western Streams and Rivers* EPS 620/R-05/005, US Environmental Protection Agency, Washington, DC.

Stoddard JL, Herlihy AT, Hill BH, Hughes RM, Kaufmann PR, Klemm DJ, Lazorchak JM, McCormick FH, Peck DV, Paulsen SG, Olsen AR, Larsen DP, Van Sickle J, Whittier TR (2006) *Mid-Atlantic Integrated Assessment (MAIA) State of the Flowing Waters Report*. EPA/620/R-06/001, US Environmental Protection Agency, Washington, DC.

Stohlgren TJ, Bull KA, Otsuki Y (1998) Comparison of rangeland vegetation sampling techniques in the central grasslands. *Journals of Range Management* 51(2) 164–172.

StreamNet (2008) *StreamNet*. Pacific States Marine Fisheries Commission. [Accessed on-line December 25, 2008 http://www streamnet org].

Svancara LK, Garton EO, Chang K-T, Scott JM, Zager P, Gratson M (2002) The inherent aggravation of aggregation: An example with elk aerial survey data. *Journal of Wildlife Management* 66(3): 776–787.

Theobald DM, Stevens DL Jr, White D, Urquhart NS, Olsen AR, Norman JB (2007) Using GIS to generate spatially balanced designs for natural resource applications. *Environmental Management* 40(1):134–146.

Thomas CE, Parresol BR (1989) Comparing basal area growth rates in repeated inventories: Simpson's Paradox in forestry. *Forest Science* 35(4): 1029–1039.

Thompson SK (2002) *Sampling*, 2nd ed. John Wiley & Sons, Inc, New York, NY.

Thompson ID, Oritz DA, Jastrebski D, Corbett D (2006) A comparison of prism plots and modified point-distance sampling to calculate tree stem density and basal area. *Northern Journal of Applied Forestry* 23(3): 218–221.

Thompson WL, White GC, Gowan C (1998) *Monitoring Vertebrate Populations* Academic Press, Inc., San Diego, CA.

USDA-FS (US Department of Agriculture – Forest Service) (1997) *Field Instructions for the Inventory of Western Oregon: 1995-97* USDA Forest Service, Pacific Northwest Research Station, Pacific Resources Inventory, Monitoring, and Evaluation Program, Portland, OR.

USDA-FS (US Department of Agriculture – Forest Service) (1998) *Field Instructions for the Inventory of Eastern Oregon: 1998* USDA Forest Service, Pacific Northwest Research Station, Pacific Resources Inventory, Monitoring, and Evaluation Program, Portland, OR.

USDA-FS (US Department of Agriculture – Forest Service) (2007) *Forest Inventory and Analysis Strategic Plan: A history of success, A dynamic future*. Forest Service Report FS-865. USDA Forest Service, Washington, D.C.

US EPA (US Environmental Protection Agency) (2002) *Research Strategy: Environmental Monitoring and Assessment Program* EPA 620/R-02/002 US Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC.

US EPA (US Environmental Protection Agency) (2006) *Wadeable Streams Assessment: A collaborative Survey of the Nation's Streams.* EPA 641 B-06/002. US Environmental Protection Agency, Washington, DC.

US EPA (US Environmental Protection Agency) (2008) *Aquatic Resource Monitoring: Monitoring Design and Analysis.* US Environmental Protection Agency, Washington, DC. [Accessed on-line March 20, 2008 http://www epa gov/nheerl/arm/designpages/design&analysis htm]

Van Sickle J, Hughes RM (2000) Classification strengths of ecoregions, catchments, and geographic clusters for aquatic vertebrates in Oregon. *Journal of the North American Benthological Society* 19(3): 370–384.

Vaux PD, Whittier TR, DeCeare G, Kurtenbach JP (2000) Evaluation of a backpack electrofishing unit for multiple lake surveys of fish assemblage structure. *North American Journal of Fisheries Management* 20(1): 168–179.

Waite IR, Herlihy AT, Larsen DP, Klemm DJ (2000) Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. *Journal of the North American Benthological Society* 19(3): 429–441.

Wagner CH (1982) Simpson's Paradox in real life. *The American Statistician* 36(1): 46–48.

Whittaker J (1984) Model interpretation from the additive elements of the likelihood function. *Applied Statistics* 33(1): 52–64.

Wiens JA (1989) Spatial scaling in ecology. *Functional Ecology* 3(4): 385–397.

Winslow SR, Sowell BF (2000) A comparison of methods to determine plant successional stages. *Journal of Range Management* 53(2): 194–198.

Wong D (1996) Aggregation effects in geo-referenced data. Pages 83–106 in: *Practical Handbook of Spatial Statistics* (ed. Arlinghaus SL) CRC Press, Inc, Boca Raton, FL.

WSDE (Washington State Department of Ecology) (2006) *Status and Trends Monitoring for Watershed Health and Salmon Recovery: Quality Assurance Monitoring Plan*. Ecology Publication No. 06-03-203. Washington State Department of Ecology, Olympia, WA.

Yule GU, MG Kendall (1950) *An Introduction to the Theory of Statistics*. Hafner Publishing Company New York, NY.

Appendix A. Hypothetical example of a lurking variable.

The examples below illustrate the operation of lurking variables. The variable in this case is geography: the groups represent studies conducted in different regions, e.g. different watersheds.

Example 1.

Figure A-1 is a scatter plot showing a response observed at different levels of some stressor; observations were collected over three disjoint spatial units, e.g. watersheds. There is no apparent strong relationship between stressor and response; a regression line fitted to the data yields a slightly negative slope, but the slope does not differ significantly from zero. Based on this evidence, one would conclude that the stressor has little or no impact on the response.

A different picture emerges if the regional origins of the data are identified. Figure A-2 shows the same data as in Figure A-1, but with different colors and symbols for each region. Separate regression lines are fitted to data from each region, and also to the stressor—response pairs of regional means. Several aspects should be noted: the slope of each regional response is positive, and significantly so; the slope of the lines fitted to all data without regard to region is essentially zero, and the slope of the line fitted to the aggregated regional means is significantly negative. These features illustrate the kinds of potential misinterpretations that can occur with spatially aggregated data.

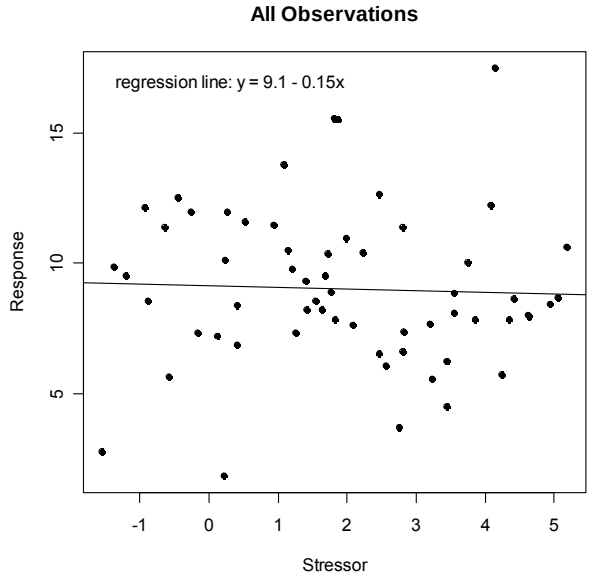

Figure A-1. Synthetic data from three regions. Regression line was fitted at all observations. Slope is slightly negative, but is not significantly different from 0 (P=0.8).

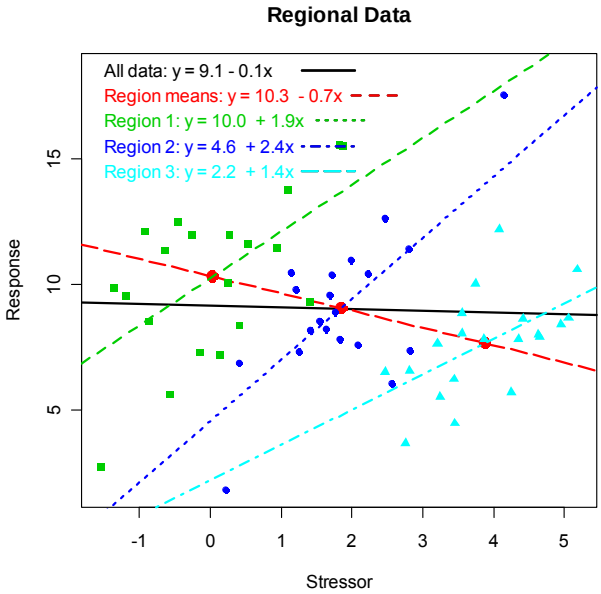

Figure A-2. Regional data with fitted regression lines. Lines were also fitted to all data, and to pairs of regional means. Note that while all three regional slopes are positive (P<0.02), the slope of the line fitted for all data is essentially zero (P = 0.8), and the regional mean slope is negative (P= 0.003).

Example 2.

Instead of a relationship being obscured by aggregation over space as illustrated above, aggregation can create an apparent relationship at a larger spatial extent where none exists at smaller spatial extents. Figure A-3 shows another synthetic data set, this time from five regions. In this case, the regression line fit to the composite data set shows a strong, highly significant negative association. As with the first example, looking at the individual regions in Figure A-4 gives a different interpretation: the individual slopes show little coherency with slopes ranging from −0.15 to +0.25.

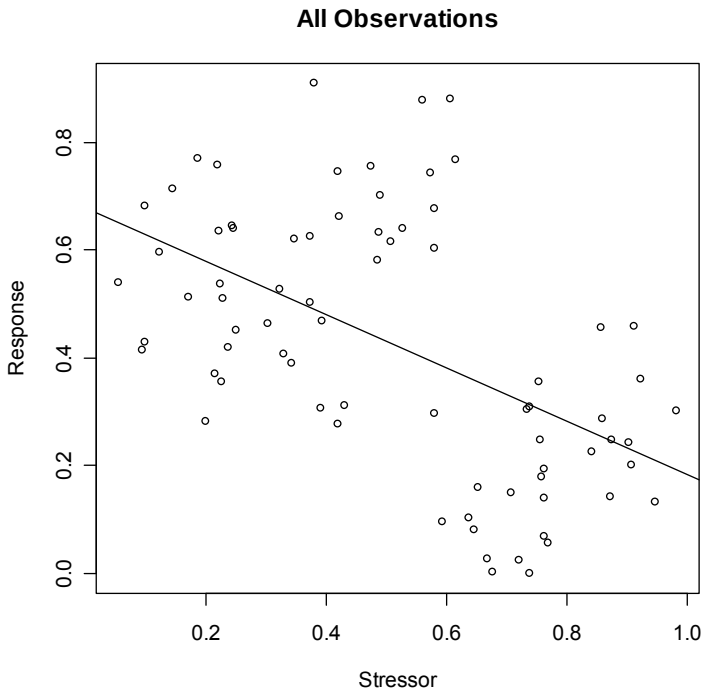

Figure A-3. Synthetic data from five regions. The slope of regression fitted to all data is negative (−0.49, P ≈ 0).

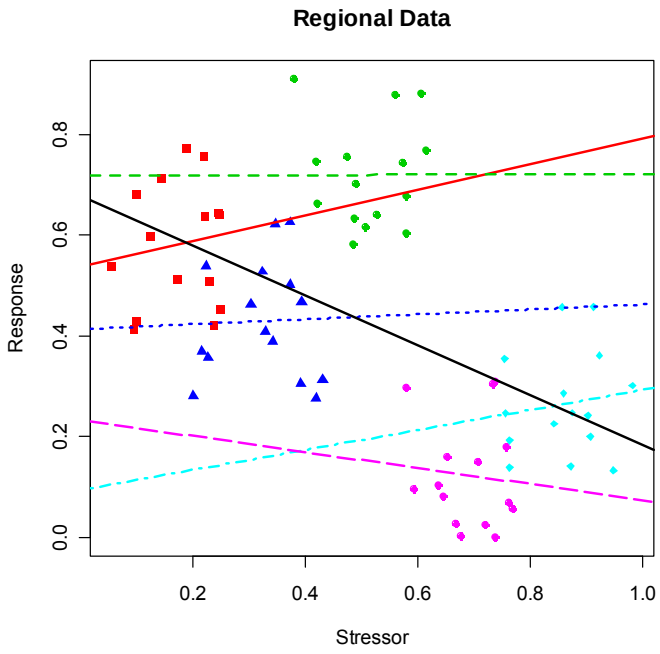

Figure A-4. Regional data with fitted regression lines. Slopes range from -0.15 up to +0.25, but none are significantly different from zero.

Studies can be designed to enhance recognition of "lurking" variables. One technique is to ensure that the same levels of the controllable treatment, or "stressor" variables, are applied over

all watersheds or geographic regions. Although there may still be a hidden geographic effect on the response, the presence of a lurking variable may be easier to discern with uniform treatment levels. Cressie (1996) describes an approach for adjusting a regression for a lurking geographic variable, and the discussion here is similar. The approach presented here is not a panacea, but does suggest an avenue that should be fruitful. Cressie (1996) suggested that one way to account for impact of spatial pattern on relationships was to include an explicit geographical variable in the relationship, e.g., in a linear regression, to consider the model $Y(s) = \beta_0 + \beta_1 X(s) + \beta_2 G(s) + \varepsilon(S)$ instead of the model $Y(s) = \beta_0 + \beta_1 X(s) + \varepsilon(S)$. The variable *G(s)* is included to explicitly account for the impact of spatial variation. Cressie arbitrarily defined *G(s)* for a geographic region to be the rank of the mean of *Y* over that region. In the examples, then, each of the watersheds or geographic subsets would have different values of *G* corresponding to ranks of the means of the *Y*. The adjustment on the response and stressor was accomplished by regressing each individually on *G* to obtain $\hat{Y} \mid G$ and $\hat{X} \mid G$, where $\hat{Y} \mid G = \hat{\beta}_0 + \hat{\beta}_2 G$, with $\hat{X} \mid G$ defined analogously. The stressor and response variables, adjusted for the effect of geography, are then the residuals $Y - \hat{Y} \mid G$ and $X - \hat{X} \mid G$. The slope of the relationship between *X* and *Y* is estimated as the slope of the regression of $Y - \hat{Y} \mid G$ on $X - \hat{X} \mid G$. The resulting coefficient is called the partial regression coefficient of *Y* on *X* after adjusting for the effect of *G*.

Using the ranks of the regional means of *Y* for the geographical variable imposes some constraints that may not be necessary or desirable. Instead, the geographical variable can be viewed as a factor with one level for each region. That approach is illustrated here by analyzing the two previous examples. The concept is the same; the relevant slope is estimated as the slope of the regression of $Y - \hat{Y} \mid G$ on $X - \hat{X} \mid G$. The results are shown in Figures A-5 and A-6. For Example 1, the result is a strong indication of a positive slope, while for Example 2, there is no evidence of a relationship between the two variables.

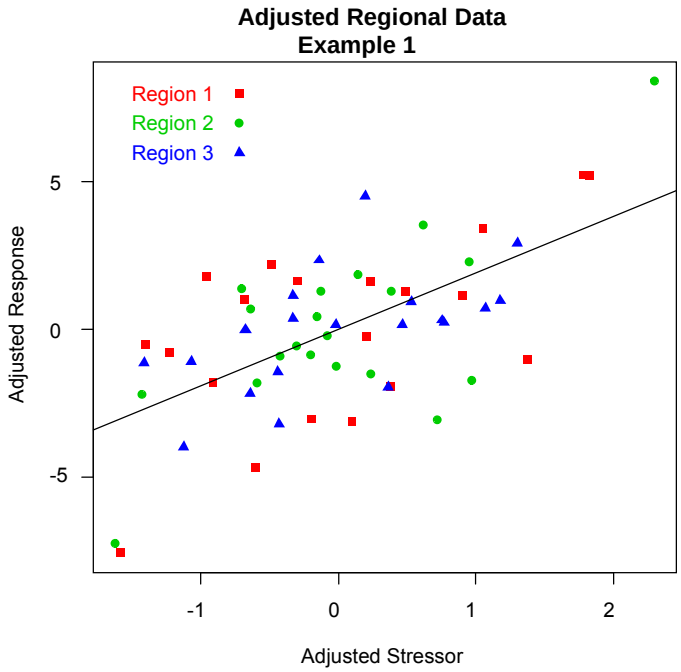

Figure A-5. Regional data from Example 1 (see Figures A-1 and A-2) adjusted for "geography", with a fitted regression line. Note that data from all regions is intermixed on both stressor and response scales, so that the data from all three regions has a coherent slope (Fitted slope is 1.9, P ≈ 0).

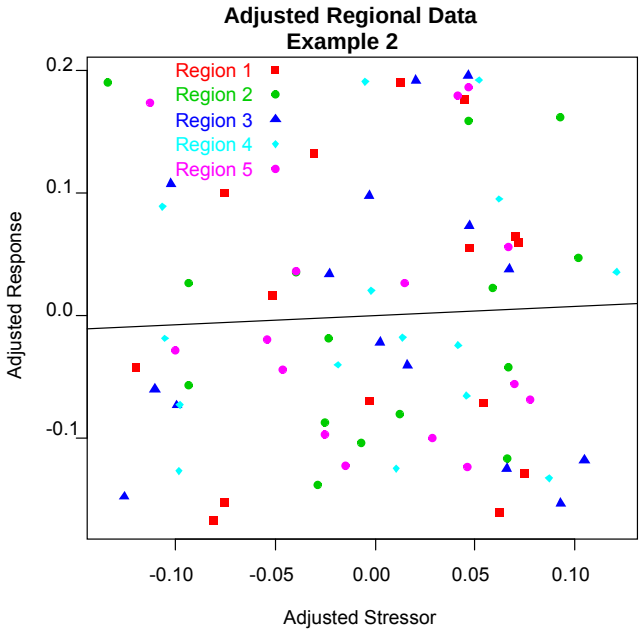

Figure A-6. Regional data from Example 2 (see Figures A-3 and A-4) adjusted for "geography", with a fitted regression line. Note that data from all five regions is intermixed on both stressor and response scales. Slope of the regression line is slightly, but not significantly, positive (P = 0.68)

It can be challenging to develop a reasonable representation of a geographical variable. In Examples 1 and 2, the geographic subsets or regions were assumed to be well-defined beforehand. That may be the case if the data were collected in identifiable regions, such as distinct watersheds. Unfortunately, that is not always the case, and there may be no obvious *a priori* regional definitions. The adjustment used in the examples can be thought of as simply re-centering the two variables for each region (i.e., subtracting the regional means from the two variables). This is a reasonable strategy if it is believed that the geographical variable is expressed as an additive impact to the stressor and response.

Appendix B. Sampling protocol and accuracy comparison studies

Table B1. Summary of published studies that compared sampling protocols.

Monitoring subject	Location	Action	Findings	Reference
Stream macroinvertebrates	Mid-Atlantic, US	Compared single and multiple habitat protocols for sampling macroinvertebrates.	Index scores were highly correlated between methods. Sampling comparability was high for species level data. Variability in relationships between the methods indicated that data were not interchangeable.	Blocksom *et al.* 20088
Stream macroinvertebrates	California	Compared 3 bioassesment methods.	Assessment scores were highly correlated (e.g., >0.839) among methods. Methods were able to distinguish reference from test (impaired) sites with similar accuracy and could possible be used interchangeably.	Herbst & Silldorf 2006
Stream classification (observer variation)	Oregon	examined observer differences in determining Rosgen stream types	Monitoring groups and their field crews often differed in stream type determination. In only 33% of streams assessed did all monitoring groups and field crews agree on the stream types. Variability would probably decrease with training.	Roper *et al.* 2008
Solar exposure (stream)	central & eastern Oregon	Compared hemispherical images to densiometer.	Metrics for both systems were strongly correlated for point and reach scales but not always in a linear fashion (e.g., R2 =0.78 for proportion of visible sky).	Ringold *et al.* 2003
Canopy cover (forest)	western Oregon	Compared 4 ground based cover estimation techniques in Douglas-fir/western hemlock forests.	Significant differences were found between the 4 methods. Differences in methods were not related to stand structure types. Regression equations were provided to allow for conversion of data between methods used in similar forest stands.	Fiala *et al.* 2006

Table B1. Summary of published studies that compared sampling protocols, *continued*

Monitoring subject	Location	Action	Findings	Reference
Canopy cover (riparian)	across Oregon	Compared 3 instruments used to measure riparian canopy cover.	Canopy cover measured with clinometer and hemispherical images were similar. Densiometer measurements were typically lower than the other 2 methods.	Kelley & Krueger 2005
Forest stand density	Ontario, Canada	Compared 2 methods used to determine tree stem density and basal area.	Trees with > 10 cm diameter at breast height (DBH) had a significant correlation between estimated basal area and stem densities derived from both measurements. Prism plots measured a significantly higher stem density than point distance technique, but no differences were seen in basal area estimations	Thompson *et al.* 2006
Vegetation (forest)	Wisconsin Michigan	Compared 3 methods for sampling understory and overstory vegetation and their power to detect change.	All methods detected changes in composite variables (e.g., species richness) but lacked statistical power to detect 20% change in abundance in most individual species. High power in determining change in overstory tree composition but differed in ability to track changes in understory composition and diversity.	Johnson *et al.* 2008
Vegetation (rangeland)	Colorado Wyoming S. Dakota Minnesota	Compared 4 rangeland vegetation sampling methods used to determine species diversity.	ANOVA found significant method and prairie type effects but no interactions between the two for total species richness, the number of native species, and species with less than 1% cover. The methods produced similar results for total foliar cover and soil cover	Stohlgren *et al.* 1998

Table B1. Summary of published studies that compared sampling protocols, *continued*

Monitoring subject	Location	Action	Findings	Reference
Vegetation (rangeland)	Montana	Compared 2 methods used to determine rangeland plant successional stages.	Range condition scores were significantly greater than ecological status scores. *Range condition analysis* and *Ecodata* produced similar condition classes (e.g., poor, good, excellent) but different numerical condition scores.	Winslow & Sowell 2000
Forage use (rangeland)	Arizona	Compared 3 techniques used to determine forage use by ungulates (cattle, elk) in rangelands.	The two height to weight methods produced lower estimates than the paired plot method. The paired plot method was more precise than either of the 2 height-weight methods.	Halstead *et al.* 2000
Acorns (mast)	Arkansas	Compared 5 types of visual mast surveys to seed trap data.	Indices derived from each visual survey method were highly correlated (r ranged 0.81–0.87) with seed trap data. Surveys using fewer than 6 subjective categories produced significantly different acorn densities among all categories, whereas surveys using 9 or 10 categories did not.	Perry & Thill 1999

Table B2. Summary table of published studies that examined the accuracy of volunteer-collected data.

Monitoring subject	Location	Method	Findings	Reference
Urban forest inventory	Massachusetts	Compared trained volunteers' data with parallel data collected by arborists.	Tree genus agreement between volunteers and professionals ranged from 91 to 96%. Genus & species agreement ranged from 46 to 96% and was particularly low for *Platanus* and *Fraxinus* species.	Bloniarz & Ryan 1996
Forest stand survey	central Washington	Compared Oregon white oak stand inventories by students (grade 3–10) to inventories conducted by professionals.	No difference found in oak diameter within each transect between both classes of student (i.e., <6-grade and ≥6-grade). Subjective crown assessments and live or dead status differed. No difference was found in the proportion of three oak crown classes between students and professionals. Differences did occur with the classification of mushroom-shaped crowns.	Galloway *et al.* 2006
Forest vegetation surveys	Illinois	Compared trained volunteers' data to parallel data collected by botanists.	No significant difference found for volunteer recorded frequencies of 12 out of 20 species. Accuracy rate for 12 (out of 15) genera was 80% or higher. Frequencies differed significantly for *Carya* species., *Morus rubra*, *Ostrya virginiana*; *Ulmus* species., and some *Quercus* species. Botanists found a greater number of tree and shrub species. Volunteer counts underestimated tree species richness by 18% and shrub species richness by 33%.	Brandon *et al.* 2003
Tropospheric ozone exposure in white pine	New Hampshire	Compared student pine needle measurements to laboratory spectral reflectance measurements made on student collected branch samples.	Student-derived data (from 1991 through 1996) was found to correlate well with spectral measurements (R^2=89).	Rock & Lauten 1996
Bees used to monitor environmental pollution	Puget Sound area, Washington	Audit conducted to determine if beekeepers' performance met quality assurance standards.	No more than a 5% variance occurred between the brood score sheets recorded by volunteers and those recorded by auditors. Procedural errors identified in the audit were corrected and the number of valid tests increased the following year.	Bromenshenk & Preston 1986

Table B2. Summary table of published studies that examined the accuracy of volunteer-collected data, *continued.*

Monitoring subject	Location	Method	Findings	Reference
Water quality – acid rain monitoring	Massachusetts	Validated volunteers' implementation of sampling procedure. Compared volunteer collected data to professional data.	Sites chosen by volunteers showed no significant bias toward either more or less acidic conditions. No significant difference found in alkalinity between professional and volunteer collections (highly correlated $r^2 = 0.986$). Significant differences were found for pH (volunteer samples showed slightly higher pH at lower pH values, and slightly lower pH at higher pH values) but the agreement was good ($r^2 = 0.949$).	Mattson *et al.* 1994
Water quality	Missouri	Compared trophic state assessments based on volunteer-collected samples with assessments based on university-collected samples.	Volunteer and university trophic state classifications were identical for 74% of total phosphorus comparisons, for 84% of total nitrogen comparisons, and 89% of algal chlorophyll samples. Split sampling showed no significant differences for total suspended solids, algal chlorophyll or total nitrogen. Volunteer total phosphorus samples were significantly lower than university samples and may have reflected differences in storage methods.	Obrecht *et al.* 1998
Water quality	Florida	Compared lake water quality data collected by volunteers to data collected by professionals.	Mean Secchi disk depth, total nitrogen, total phosphorus, and chlorophyll values were strongly correlated ($r > 0.99$) to the mean values obtained by professionals.	Canfield *et al.* 2002
Water quality	Australia	Compared volunteer protocol and data collected to professional protocol and data collected.	Three of 23 annual datasets recoded showed statistically different results between volunteer and professional data for turbidity ($P<0.05$). Volunteer turbidity data appeared to be least accurate at the highest and lowest turbidity levels. No differences were seen in electrical conductivity or pH; however, volunteer pH data had higher variability than professional pH data.	Nicholson *et al.* 2002

Table B2. Summary table of published studies that examined the accuracy of volunteer-collected data, c*ontinued.*

Monitoring subject	Location	Method	Findings	Reference
Water quality	Oregon	Assessed *E. coli* quality control data that resulted from side-by-side samples collected by volunteers and agency staff.	95% of agency samples fell within 0.6 log units of each other. 97% volunteer-vs.-agency side-by-side samples fell within 0.6 log units of one another.	Hanson 2006
Benthic macroinvertebrates	Puget Sound area, Washington	Compared volunteer collected sampling and data to professional sampling and collected data. Both groups used the same protocols and sites .	A summary multimetric index was used for the comparison. No significant difference found between field samples collected by volunteers and professionals. The ability of the index to detect significant differences among sites (statistical power) improved by only 13% for assessments based on professional lab identification instead of volunteer lab identification.	Fore *et al.* 2001
Benthic macroinvertebrates	Virginia	Compared biological assessments by volunteers to those by professionals. Protocols were modified and revalidated	In the first assessment the numerical results from volunteers did not correlate well with professional samples (r = 64) and at times produced different conclusions on ecological condition (65% agreement). Volunteer protocol consistently overrated ecological condition. Protocols were modified and a second assessment was done; the new volunteer multimetric index correlated well with a professional index (r=0.6923). Conclusions about ecological condition reached by volunteers and professional protocols agreed closely (96%).	Engel & Voshell 2002
Benthic macroinvertebrates	Minnesota	Examined possible sources of volunteer bias associated with organism identification by comparing the relative size and movement of organism in volunteer vs. professional samples. Study used <u>untrained</u> volunteers.	Untrained volunteers showed a bias toward selecting larger organisms and those with more movement. The likelihood of correct family-level identification was 1.8 times greater if the family was present on a reference card. Mean success rate for identifying organisms using a family-level key was 29.6%.	Nerbonne & Vondracek 2003

Table B2. Summary table of published studies that examined the accuracy of volunteer-collected data, *continued.*

Monitoring subject	Location	Method	Findings	Reference
Spiny water flea (*Bythotrephes longimanus*) lake invasions	Ontario, Canada	Determined a volunteer program's ability to detect invasions. Professionals used volunteer methods.	Sampling methods and nets used were sufficient to detect the presence of the organism. In one lake 3 sampling stations were sufficient but in second lake three stations were not enough – seven were needed. The volunteer program failed to detect invasion 14% of the time.	Boudreau & Yan 2004
Amphibian call surveys	Michigan	Evaluated the affect of volunteer observer experience on data quality.	Volunteers were relatively reliable in their abilities to identify most species by call, but there was extensive variability in abundance estimation. Some species were characteristically confused (e.g., *Rana pipiens* and *R. palustris*) and additional species were frequently recorded as present when they were actually absent.	Genet & Sargent 2003
Marine reef fish census	Mafia Island, Tanzania	Tested and validated use of volunteer divers to census reef fish populations.	Volunteers' ability to indentify 56 species in 30 genera increased significantly between two censuses. After 11 additional dives the loss of precision compared with an experienced "control" diver was reduced from 13% to 0.6%.	Darwall & Dulvy 1996
Marine fish survey	Florida	Evaluated a standardized visual survey method designed for volunteer SCUBA divers. Results were compared to 2 independent quantitative studies from 1978 and 1994.	Experienced divers were able to provide useful species listings, frequency of occurrence and abundance data.	Schmitt & Sullivan 1996
Marine mollusk used to monitor environmental pollution	Great Britain	Compared data collected by volunteers on imposex (penis development on marine gastropods) in dogwhelk (*Nucella lapillus*) to data collected by a professional.	Data collected by volunteers were correlated closely with data collected by a professional (measures for one index r = 0.699 and for a second index r=0.865). Volunteers tended to produce higher assessments for both indices than the professional.	Evans *et al.* 2000

Table B2. Summary table of published studies that examined the accuracy of volunteer-collected data, *continued.*

Monitoring subject	Location	Method	Findings	Reference
Marine biodiversity – invertebrate and seaweed species	Great Britain	Compared distribution and abundance data on littoral animals and plants collected by volunteers to data collected by professionals.	For 6 of 8 species, there was >80% agreement between volunteer and professional abundance data. Volunteers tended to assess each invertebrate species more generously (i.e. used abundance categories such as super abundant, abundant, and common, more often) than scientists.	Evans *et al.* 2001
Marine biodiversity – invertebrate and seaweed species	Scotland	Compared distribution and abundance data on littoral animals and plants and length measurements of gastropods collected by volunteers to data collected by professionals.	Abundance assessments by volunteers varied considerably for 3 of 4 target species. Volunteers tended to interpret subjective abundance categories (e.g., super abundant) differently from one another. Volunteer data was sufficiently robust to produce reliable abundance maps for only 1 on the 4 target species.	Foster-Smith & Evans 2003
Marine conservation survey - seahorses	Italy	Assessed volunteer performance in survey. No professional survey was done for comparison.	Seahorse identification was not difficult because there were clear morphological differences between the two species. Data were consistent across 3 years. The greatest limitation with volunteers as identified by authors was the difficulty in obtaining a uniformly sample across time and space.	Goffredo *et al.* 2004
Warm-water game fish	Mississippi	Compared angler diary catch rates to creel survey and electrofishing catch rates. Compared fish population length distributions obtained from angler diary to electrofishing surveys.	No significant ($P<0.05$) correlations were found between angler diary catch per unit effort and creel survey or electrofishing catch per unit effort. Length distributions of black bass obtained from diaries and electrofishing were similar at 5 of 7 reservoirs but distributions were different for crappies.	Bray & Schramm 2001

Table B2. Summary table of published studies that examined the accuracy of volunteer-collected data, *continued*.

Monitoring subject	Location	Method	Findings	Reference
Bird song/call surveys	Ontario, Canada	Tested if volunteers with low to moderate skill levels could learn to identify and count forest birds by song or call. Compared inexperienced volunteers to experienced volunteers.	No difference was found in either counts of individual species (12-13 target species) or in the suite of species present between experienced-experienced pairs and experienced-novice pairs. Novices tended to count fewer birds with significant differences for 3 species in 1995 and 1 species in 1996.	McLaren & Cadman 1999