

Issues in the Aggregation of Data to Assess Environmental Conditions

Executive Summary

An Excerpt from:

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.

Full report available at <http://www.fsl.orst.edu/imst>

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.