

Measurement and Meaningfulness in Conservation Science

Key words: Biodiversity value, estimated data, expert judgement, measurement scale, systematic conservation planning.

Abel G. Wolman

Center for Large Landscape Conservation

and

AGW Consulting, Inc.

855 NW Lincoln Street

White Salmon, WA 98672-4326

wolman@gorge.net

Abstract

Incomplete databases often require conservation scientists to estimate data either through expert judgment or other scoring, rating, and ranking procedures. At the same time, ecosystem complexity has led to the use of increasingly sophisticated algorithms and mathematical models to aid in conservation theorizing, planning, and decision making. Understanding the limitations imposed by the scales of measurement of conservation data is important for the development of sound conservation theory and policy. In particular, biodiversity valuation methods, systematic conservation planning algorithms, geographic information systems (GIS), and other conservation metrics and decision-support tools, when improperly applied to estimated data, may lead to conclusions based on numerical artifact rather than empirical evidence. The representational theory of measurement is described here, and the description includes definitions of the key concepts of scale, scale type, and meaningfulness. Representational measurement is the view that measurement entails the faithful assignment of numbers to empirical entities. These assignments form scales that are organized into a hierarchy of scale types. A statement involving scales is meaningful if its truth value is invariant under changes of scale within scale type. I apply these concepts to three examples of measurement practice in the conservation literature. The results of my analysis suggest that conservation scientists do not always investigate the scale type of estimated data and hence may derive results that are not meaningful. Recognizing the complexity of observation and measurement in conservation biology, and the constraints that measurement theory imposes, the examples are accompanied by suggestions for informal estimation of the scale type of conservation data and for conducting meaningful analysis and synthesis of this information.

Introduction

Conservation scientists study complex systems for which there may be inadequate or incomplete data (Possingham et al. 2000; Groves 2003; Wilson et al. 2005). They also face the self-imposed task of applying what they have learned “to help transform the knowledge we generate into sound public policy” (Meffe 1998). The problem of incomplete conservation data sets has produced a variety of data estimation techniques including expert judgement and various scoring, rating, and ranking procedures (Groves 2003). The challenge of creating sound public policy has led to calls for systematic approaches to conservation planning (Pressey 1994; Margules & Pressey 2000) with increased emphasis on algorithmic and mathematical tools for efficient reserve selection and design of wildlife corridors and conservation areas (Williams et al. 2004). Meir et al. (2004) noted that “most large conservation organizations and natural resource agencies now use systematic conservation planning methods to identify optimal or near optimal reserve networks.”

The analysis of estimated data by automated decision-support tools is both beguiling and risky. It is beguiling because data can be generated and analyzed in situations where previously little was known and risky because the information gleaned from such efforts may be artificial—the result of numerical relationships that have no underlying empirical counterparts. The problem of meaningful analysis of estimated data is not, however, restricted to high tech applications and may be illustrated by a simple example. Suppose two potential reserve sites X and Y received expert estimated habitat suitability scores of 2 and 4, respectively (5, say, being the highest possible score, 1 the lowest). During public hearings, a policy maker asserts that site Y has twice the suitability of site X . What is the empirical status of this statement? In particular, is it meaningful? That is, do the facts in the field warrant the policy maker’s arithmetic? Intuitively, one may sense that expert derived suitability scores do not justify the policy maker’s conclusion. So how can this intuition be formalized, and how does this formalization pertain to automated and non-automated approaches to conservation planning?

I describe measurement-theoretic limitations to conclusions in conservation science and planning based on information from different measurement scales. I define the representational

measurement theory concepts of scale type and meaningfulness and relate these to the problem of analyzing estimated conservation data. I argue that the concept of meaningfulness formalizes our intuition concerning the above apocryphal policy maker's arithmetic, and is essential for understanding and limiting the introduction of numerical artifact into the analysis and synthesis of estimated data.

Nearly 20 years ago, Smith and Theberge (1987) described the application of measurement theory to conservation evaluation. They were surprised that conservation evaluation had, "for the most part, ignored the knowledge gained in...other disciplines in coping with...[measurement-theoretic] issues." These issues pertain to the everyday activity (in conservation science and elsewhere) of assigning numbers to things and then doing arithmetic on these numbers as if they represented the same sort of measurable quantities as lengths or weights. Given the obvious difficulty of measuring mental events, psychology was one of the first fields to grapple with these issues. The problems now intrude on other areas of inquiry including conservation. Smith and Theberge's discussion of measurement theory did not resonate with conservation scientists (but, see Chrisman 1998, 2002; Pressey & Tuffs 2001). Here, I expand on the work of Smith and Theberge with the hope of convincing conservation scientists and planners that measurement-theoretic ideas deserve further attention.

Representational Measurement Theory

The basic premise of representational measurement theory is that measurement consists in the assignment of numbers to empirical entities. These assignments are called *representations* or measurement *scales*. If one thinks of the empirical entities as forming a set E , then a scale can be viewed as a function from E into the real numbers. The fundamental requirement of representational measurement is that the process of assigning numbers to elements of E should preserve the empirical relations on E . This means, in particular, that different empirical systems may involve different types of scales. It also means that inferences or conclusions based on measurements of an empirical system should reflect intrinsic properties of the system and not

artifacts of the numerical representation. The psychophysicist S. S. Stevens (1946) gave the first mathematical formulation of these requirements introducing the central notions of admissible transformation, scale type, and meaningfulness. In the following, I describe an informal or non-axiomatic approach to representational measurement based on modern refinements of Stevens' original ideas. A formal, axiomatic treatment of representational measurement is given in the three volume work by Krantz et al. (1971), Suppes et al. (1989), and Luce et al. (1990). Roberts (1979) offers a more applied reference.

To motivate the definitions of admissible transformation and scale type, I consider the well-known conversion between Celsius and Fahrenheit temperature scales. The conversion formula is given by the linear function $\phi(x) = (9/5)x + 32$. If one thinks of the Celsius scale C as a measurement in the representational sense defined above, that is, as a function which assigns to an object or entity e a real number $C(e)$, its temperature in degrees Celsius, then the school-taught method for converting from Celsius to Fahrenheit can be equivalently understood as applying the conversion function ϕ to the values or output of the Celsius scale C . The process of applying a function ϕ to the output of another function C is called the composition of ϕ and C and is denoted by $\phi \circ C$. The conversion from Celsius to Fahrenheit scales is now compactly expressed by the equation $F = \phi \circ C$. (For functions f and g one defines a new function $h = f \circ g$ called the composite of the functions f and g by setting $h(x) = (f \circ g)(x) = f(g(x))$. The right hand side of this equation expresses the rule "apply f to the output ($g(x)$) of g .")

Stevens (1946) called scale conversion functions like ϕ *admissible transformations* because they transform one acceptable measurement scale into another. He further recognized that measurement scales could be classified according to the types of admissible transformations that convert between them. For example, any acceptable temperature scale f can be converted into another acceptable temperature scale g by composing f with an appropriate linear function ϕ . Temperature scales are thereby characterized by the class of linear functions that transform one temperature scale into another.

An *admissible transformation* is defined as a function that changes one acceptable measurement scale into another via composition of functions. More precisely, given acceptable measurement scales f and g , an admissible transformation is a function ϕ such that $g = \phi \circ f$. To certain classes of admissible transformations one associates certain classes of scales called *scale types*. Specifically, a scale type is a class of scales that can be converted amongst themselves by composition with elements from a distinguished class of admissible transformations. The following is an abbreviated taxonomy of scale types. Table 1 summarizes these examples.

| Admissible transformations | Scale type | Example |
|---|------------|---|
| $\phi(x) = x$ (identity) | absolute | relative frequency counting |
| $\phi(x) = \alpha x, \alpha > 0$ | ratio | weight (mass) length temperature (kelvins) time (duration) |
| $\phi(x) = \alpha x + \beta, \alpha > 0$ | interval | temperature (Fahrenheit, etc.) time (calendar) |
| $x \geq y$ iff $\phi(x) \geq \phi(y)$ (monotone) | ordinal | Mohs hardness scale rankings grades (school) |
| ϕ one-to-one (permutation) | nominal | naming classification (species) |

Table 1. Admissible transformations, their associated scale types, and example scales for each scale type.

If no aspect of a measurement scale can be varied—that is, one is not allowed to change the unit of measurement of the scale or anything else—then the class of admissible transformations for this scale consists of a single transformation, the identity function $\phi(x) = x$. Scales whose only admissible transformation is the identity are said to have *absolute* scale type. Examples are relative frequency and counting.

If the unit of measurement of a scale is allowed to vary, but nothing else, then the class of admissible transformations consists of *similarities* $\phi(x) = \alpha x$ with real constant $\alpha > 0$. Scales

related by a simple change of unit are of *ratio* scale type. Weight (mass) measurements are ratio scales (short for “scales of ratio scale type”). An example of an admissible transformation for weight is the function $\phi(x) = (2.2)x$ converting kilograms to pounds. Other ratio scales are length, time duration, and absolute temperature (measured in kelvins).

If both the unit of measurement and the zero point are allowed to change, then the class of admissible transformations are linear functions of the form $\phi(x) = \alpha x + \beta$ for real numbers $\alpha > 0$ and β . Scales related by linear transformation are called *interval* scales. As discussed above, temperature scales are interval as is the measurement of time by calendar.

When only the ordering of the measurements matters (i.e., when one can say something is bigger or smaller than something else, but cannot say how much bigger or smaller it is), then the admissible transformations are monotone or increasing functions ϕ satisfying $\phi(x) \leq \phi(y)$ if and only if $x \leq y$. The associated scales are said to have *ordinal* scale type. Examples include Mohs hardness scale and various grading, ranking, and rating procedures.

Finally, if the class of admissible transformations consists of all one-to-one functions (permutations in the finite case), then the resulting scales have *nominal* scale type. Nominal scales encode naming or classification of objects or events.

The narrower the class of admissible transformations the higher or stronger the measurement scale. Absolute scales are the strongest scale types, nominal the weakest. The common distinction between qualitative and quantitative data is marked by the transition from ordinal to interval scales.

Representational measurement is based on the idea of numerically *representing* an empirical system. This implies that the truth or falsity of results derived from measurements should not depend on a fortuitous choice of scale. Consider the statement: “It is presently 60° F in Baltimore and 30° F in Bozeman, so it is twice as warm in Baltimore as in Bozeman.” What is the status of this statement? Is it true or false? In fact, its validity is ambiguous because it is scale dependent—in degrees Fahrenheit it is true; in degrees Celsius, false. To guard against this sort of ambiguous statement representational measurement applies the concept of

meaningfulness: A statement involving numerical scales is *meaningful* if and only if its truth or falsity is unchanged after all of the scales in the statement are transformed by any admissible transformation (Roberts 1979). A statement that is not meaningful is said to be *meaningless*.

The statement involving temperatures is meaningless because after converting from Fahrenheit to Celsius, that is, after applying an admissible transformation $\phi(x) = (5/9)x - 160/9$ to the interval temperature scales, the statement becomes false: $-40/9 \neq 140/9$. Although this example may seem inconsequential, it actually implies something noteworthy: It is meaningless in general to assert that one interval scale value is a multiple of another.

Meaningfulness is different from truth. The statement “I am 10 times taller than the Empire State Building” is meaningful, but false. Also, in everyday usage, *meaningful* and *meaningless* are loaded terms. They are used here in their measurement-theoretic sense only.

Meaningfulness Examples

I present several idealized applications of the concept of meaningfulness. Although simple, these initial cases are important. The calculations relate directly to the meaningfulness of conservation modeling and decision-support methodologies.

Consider, again, the policy maker’s statement, “site X has twice the suitability of site Y .” The meaningfulness of this statement depends on the scale type of expert-assigned habitat suitability scores. Suppose suitability scores form a ratio scale. (This means experts can assess suitability in the same way and with the same proficiency that one can measure length or weight. In other words, experts have a mental “suitability ruler.”) Let S denote this scale so that

$$S(X) = 2S(Y)$$

mathematically expresses the statement: Site X has twice the habitat suitability of site Y . To check whether this statement is meaningful for ratio scales, one must show that the truth or falsity of the statement does not depend on a particular choice of unit. Recall that the admissible transformations for ratio scales are simply changes of unit $\phi(x) = \alpha x$ for $\alpha > 0$. It follows that the policy maker’s statement is meaningful if and only if the equation

$$\phi[S(X)] = 2\phi[S(Y)],$$

holds for all changes of unit ϕ . This equation says that after a change of unit, that is, after we have applied the admissible transformation ϕ , it is still the case that site X has twice the habitat suitability of site Y . Applying the definition of ϕ yields, equivalently, that the equation

$$\alpha \cdot S(X) = 2\alpha \cdot S(Y)$$

must be true for all $\alpha > 0$. Because this is the case given the truth of the original assertion, it follows that the policy maker's statement is meaningful when suitability scores are ratio scales.

Suppose, instead, that expert suitability scores form an interval scale. (Experts possess a kind of "suitability thermometer.") As demonstrated in section 2, for interval temperature measurements the policy maker's statement is now meaningless. Since interval scales are a special case of ordinal and nominal scales, the statement is also false for these weaker scale types. Meaninglessness formalizes our intuitive skepticism for arithmetic results derived from qualitative, subjective scores.

A second example, relates to merging-functions for expert judgements. Suppose three experts assign habitat suitability scores $(2,2,2)$ and $(1,1,3)$ to sites X and Y , respectively. Assume the scores are assigned on an ascending 1 to 5 scale with 5 indicating the highest suitability level. Is it meaningful to assert that X has higher average suitability than Y ($2 > 1.67$)? Again, the answer depends on the scale type of the suitability scores. For (shared or common) interval or higher scales it can be shown that the statement is meaningful. However, suppose the expert judgements are ranks—an ordinal scale—and consider the monotone function ϕ defined by $\phi(1) = 1.6$, $\phi(2) = 2$, and $\phi(3) = 3$. This is an admissible transformation for ordinal scales which when applied to the original expert ranks yields the equally acceptable site scores $(2,2,2)$ and $(1.6,1.6,3)$. With respect to these new values, site X now has lower average suitability than Y ($2 < 2.061$). Because the truth value of the original statement changed after applying the admissible transformation ϕ , the statement is meaningless.

If experts do not share a common unit of measurement, then order comparisons of averages are meaningless even for ratio scales. To see this, suppose that the above suitability

scores are independent ratio scales; that is, each expert may use a different unit to measure habitat suitability. (Their mental “suitability rulers” are marked off with different units.) Let

$\phi_i(x) = \alpha_i x$ denote the ratio scale admissible transformation (change of unit) for expert $i = 1, 2, 3$, and set $\alpha_1 = \alpha_2 = 1$ and $\alpha_3 = 3$. Computing the means of the scores after admissibly changing units yields

$$(2 + 2 + 6)/3 = 3.333 < 3.667 = (1 + 1 + 9)/3.$$

The order of the averages is reversed; therefore the statement is meaningless.

These last two examples illustrate what appears to be a not uncommon practice in conservation science. Johnson and Gillingham (2004), for instance, use the arithmetic mean to aggregate ratings of ecological units contained in single GIS polygons. Unless these ratings form a common quantitative scale (i.e. interval or higher), the resulting maps may display meaningless information. When critical conservation decisions are at stake, such maps should be approached with caution.

Meaningfulness in Conservation Science

I examine real-world examples of conservation measurement in light of the results from the previous sections: systematic reserve selection for the Columbia Plateau (Davis et al. 1999), priority-setting for land birds in Canada (Dunn et al. 1999), and a reserve selection methodology based on biodiversity value (Arponen et al. 2005). These studies were selected because they appear to typify measurement practice in conservation science and because they allow reasonably transparent application of the concepts of scale type and meaningfulness. My primary findings are that (1) none of these examples discuss the scale type of their data, (2) in each example estimated data sets may contain ordinal-scale measurements, and (3) each of the studies may contain meaningless results.

I include actual examples because the general statements and cautions made thus far cannot adequately convey the possible significance of issues of measurement and meaningfulness to conservation biology, or, in particular, their potential impact on the policy aspirations of the

field. Although my critique is focused primarily on measurement methodology, meaningless manipulation of conservation data will ultimately affect the substance of these important inquiries as well.

The first example involves The Nature Conservancy's (TNC) application of systematic reserve selection to the Columbia Plateau ecoregion (Davis et al. 1999). "The exercise required integrating data on species, plant communities, land ownership and other socioeconomic factors, and combined expert opinion with computer-aided site selection modeling." Of the seven steps in the TNC's planning process, I consider three: step 4, assigning of suitability scores; step 5, automated selection of planning units; and step 6, assessment and visualization of optimal sets of planning units from step 5.

In step 4, each planning unit was assigned "a score to indicate its 'suitability' for conservation management...." This overall suitability index was defined as the weighted sum of seven factors in three categories. The category "habitat condition," contained four of these factors: "(1) Roadedness as percentage of area affected by roads. (2) Human population density by number of people per km. (3) Habitat quality based on expert opinion provided during six workshops. (4) Aquatic integrity index." Step 5 entailed the selection of planning units as potential biodiversity management areas. This involved the "relatively complex" task of selecting new conservation areas from "hundreds to thousands" of planning units. The decision problem was modeled as an integer program with the objective of optimizing "the selection of new conservation areas that collectively satisfy target representation levels for each vulnerable element." The model produced a set of planning units that best balanced "efficiency (least area) and suitability (best quality or most manageable areas)." Model output were visualized with GIS for evaluation in step 6. It was observed that this evaluation could lead to "changes in the specifications of goals or objectives or to refinement of the tentative plan..." and thereby iteration of steps 1 through 6.

Davis et al. conclude that the Columbia Plateau planning exercise is "a possible prototype for the other 62 ecoregions of the United States." They claim that

An important aspect of the exercise was a consistent and explicit representation of expert knowledge. In spite of the large size and biological complexity of the region and the large number of elements and factors that were considered, TNC was able to synthesize expert scientific knowledge and existing geospatial data.

The authors also point out that the planning exercise had “several deficiencies” including “potential bias and uncertainties caused by data quality and expert knowledge.”

Because Davis et al. do not discuss the scale type of the expert-determined habitat quality scores, the claim that the exercise produced a “consistent and explicit representation of expert knowledge” is difficult to evaluate. An unrecognized deficiency “caused by data quality and expert knowledge” is uncertainty about the measurement scale of expert opinion. If the scale type of habitat quality scores has not been determined, then it is unknown whether or not the TNC was able to *meaningfully* synthesize expert opinion and geospatial data.

The problem of synthesizing expert opinion occurs at least twice in the TNC planning process; first, when a consensus judgement must be produced during the six workshops, and, later, when the resulting habitat quality scores are submitted to the optimization process. Because we have no information about how the workshops produced a consensus expert judgement, it is pointless to speculate as to the meaningfulness of the outcome. (See Cook & Kress [1992] for a discussion of the intricacies of determining consensus judgements.) Nevertheless, it is recognized that certain optimization procedures, including linear and integer programming, may produce meaningless results when applied to interval and ordinal scales (Pekeč 1996). The greedy algorithm, on the other hand, may be meaningfully applied to ordinal data (Pekeč 1998). Hence understanding the scale type of habitat quality scores may be critical for choosing the appropriate optimization routine or heuristic for selecting meaningfully optimal sets of planning units. Without this understanding, vulnerable planning units may be selected based on numerical artifact and not with respect to an empirically justified balance of efficiency and quality.

The second example concerns a “priority-setting tool applied to Canada’s landbirds” (Dunn et al. 1999). The goal of this study was to develop a “ranking system to help set priorities for landbird species.” The ranking system was based on two complementary species lists

compiled from “concern” and “responsibility” scores. As described by the authors, the method adopted by Partners in Flight (PIF) Canada

Produces two complementary lists of species. The first, in common with most other ranking systems, indicates ‘concern’ and gives high rank to species with a restricted range, low numbers, or documented recent declines. The second list indicates ‘responsibility,’ meaning responsibility for stewardship or regionally characteristic species.

Concern and responsibility scores were defined as composites of still other indices; here I concentrate on the definition of concern.

“Two aspects of concern were considered in the overall concern score: vulnerability of species to localized threat...and population decline.” Vulnerability is a composite score defined as the average of “worldwide abundance, breadth of breeding range, and breadth of wintering range.” Table 1. in Dunn et al. lists these criteria and the categories or scoring divisions within each criterion. Global or worldwide abundance is based on the TNC National Heritage Program ranking system and is defined as follows (listed by division, TNC rank, and PIF-Canada score): abundant, G5, 1; common, G4, 2; uncommon, G3, 3; rare to uncommon, G2, 4; and very rare, G5, 5. Breadth of breeding and wintering range criteria are defined by division, percentage area, and score: very widespread, > 75%, 1; widespread, > 50 – 75% ,2; intermediate, > 25 – 50% ,3; local, > 10 – 25% , 4; very local, ≤10% , 5; and absent, missing. Population decline or national population trend is based on fairly lengthy criteria which I will not list in their entirety. For our purposes, it is sufficient to note that population decline is scored on a 1 to 5 scale (with missings), and to provide the definitions of the highest two scores. A score of 5 is assigned if there is a “statistically significant decrease of $\geq 3\%/year$ ” or a “well-documented very large decrease without count data.” A score of 4 is assigned if there is a “statistically significant decrease of 1 to 3%/year ,” a “nonsignificant decrease of $\geq 3\%/year$,” a “well-documented modest decrease without count data,” or a “poorly documented major decrease.” In scoring population decline, best available monitoring information was used, and “expert opinion was relied on if no other

information was available.” Finally, overall concern was defined as the “average of scores for vulnerability and national population trend.”

Dunn et al. (1999) report that raw data, scores, and documentation were stored in an electronic database. Fractional values for concern and responsibility scores

were rounded to whole numbers from 1 to 5 that corresponded to priority categories very low, low, medium, high, and very high, respectively...Species were grouped into habitat and migration-distance categories...Tukey Studentized Range tests were used to detect differences in mean concern scores among habitat and migration-distance categories...Digitized maps were imported into ArcInfo geographical information system...Overlay maps were then generated for species with high scores for concern and responsibility...there were no significant differences ($p < 0.05$) in mean concern scores among habitat or migration-distance categories.

Dunn et al. do not characterize the scale types of the criteria defining vulnerability and population decline scores. However, the consistency of the concern index depends on the meaningfulness of the arithmetic mean on these criteria. It is important then to determine (or at least estimate) the measurement scales of TNC abundance rankings, breeding and wintering range values, and population decline scores.

Determining scale type is a difficult problem. As Roberts (1994) observes, the place where the analysis of meaningfulness is “most vulnerable is in the need to know what kind of scale you have....” In general, “we must use the principle that the admissible transformations are those which preserve the information carried by the scale...it is often a matter of empirical judgement to determine the admissible transformations and hence the scale type.” To help make these empirical judgements, Knapp (1990) suggests asking the following types of questions:

Do you have anything for your *raw score* scale that even remotely resembles an actual unit of measurement?...[I]f a subject obtains a score of 3, you should be able to say 3 *what*. Does the scale have a zero point however arbitrary it may be? What transformations, if any, of your scale are acceptable? All order preserving transformations? Only linear transformations? No transformations whatsoever?

Turning to the PIF-Canada abundance criterion, we can ask the following questions: Is there a unit of global abundance? If a species has an abundance level of 3, then, “3 what?” Does abundance have a zero point indicating the complete absence of abundance? Does it make sense to state that “species A has 3 times the abundance of species B?” Or that the difference in

abundance between species *A* and *B* is 3? Is it acceptable to transform abundance scores by a linear function? Are we comfortable applying arbitrary monotone transformations to abundance scores?

Although I do not have enough information to undertake a systematic review of these types of questions for concern score attributes, a simple analysis suggests the following observations. First, it appears TNC abundance rankings form an ordinal scale. They possess neither a unit of measurement (an abundance rank of 3 is 3 what?), nor a zero point. Furthermore, there seems to be little empirical justification for asserting one abundance score is “3 times another” or “is 3 abundance units greater than another.” Nevertheless, there is a notion of order; hence, abundance scores may be ordinal. Second, understanding the scale type of breadth of breeding and winter range scores is not straightforward. Although based on a ratio scale (ratios of areas), the scores are defined by a nonlinear transformation of this scale and, therefore, are not an equivalent, admissibly transformed, ratio scale; that is, they do not correspond to a simple change of unit. Ascertaining the scale type of breadth of breeding/wintering range will require more careful study. It may make sense to use the area ratios themselves. Third, population decline although initially based on counts and hence apparently an absolute scale is, like breeding range, the result of a nonlinear (and non-identity) transformation. Unlike counts, the transformed 1-5 scale is bounded and has no numerical zero point implying it is also not a ratio scale. Furthermore, incomplete information about population counts may require expert-assigned population decline scores. The use of expert estimates can result in an ordinal population decline scale.

If one or more of the components of the vulnerability score are ordinal measurements, then the definition of vulnerability (and, therefore, concern) is based on a meaningless synthesis: the arithmetic mean of ordinal scores. (Actually, a synthesis of intermixed scale types may be involved. I ignore this complication here.) In this case, Tukey Studentized range tests and reports of no significant differences in mean concern scores among habitat or migration-distance categories may contain little useful information. The GIS overlays of ordinal concern values

present a further difficulty. Choropleth or isopleth maps containing equal interval (or any other) color gradients are inherently misleading given the true plasticity of ordinal scales.

Dunn et al. remind us that “no scoring system will give the ‘right’ answer for every species or every user of the system.” It is also important to understand the limitations measurement scales can impose on conclusions derived from any scoring system.

The final example concerns valuing biodiversity in reserve selection. Arponen et al. (2005) discuss “quantitative methods” for maximizing “the biodiversity value of reserve networks.” Underscoring the shortcomings of complementarity-based reserve-selection methodologies in which species are “often considered equal,” they “introduce a framework for reserve selection that includes species weights and benefit functions for [species] under- and over-representation.”

Arponen et al. (2005) point out that “it has long been recognized that species are not of equal value and should be prioritized (weighted) for conservation purposes...” Early “reserve-selection methodologies scored sites by weighting their biological (or other) features,” but ignored complementarity. On the other hand, recent algorithms emphasize complementarity and assume equal species value. To rectify this situation, Arponen et al. proposed maximizing biodiversity value for reserve selection. Specifically, they defined an optimization model

$$\max F(X) = \sum_j V_j [R_j(X), T_j, w_j]$$

subject to the total cost constraint

$$\sum_i c_i x_i < C .$$

Here $X = (x_i)$ is a selection vector with $x_i = 1, 0$ depending on whether site i is or is not included in the solution. The V_j express the value of species j and are functions of representation $R_j(X)$, species weights w_j , and targets T_j . The “important part” of the above formulation is the definition of the values V_j :

$$V_j(X) = w_j f_j [R_j(X)],$$

where w_j and f_j are the weight and benefit function, respectively, for species j . Apronen et al. (2005) suggest a number of different forms of benefit function based ideally on species biology. In the following, I concentrate on meaningfulness issues related to species weights w_j .

Apronen et al. (2005) assign species weights according to the formula

$$w_j = (pW_j^R + (1-p)W_j^N)W_j^T$$

where W_j^R , W_j^N and W_j^T represent, respectively, regional rarity, national rarity, and taxonomic value for species j . The parameter $p \in [0,1]$ specifies the balance of rarity weighting. The following is an example of the scoring of these component weights:

Our regional rarity weights were based on the *Field Flora of Finland*...and varied from 1 to 4 depending on the regions in which the species was classified as rare. For national rarity we used red-list classifications...ranging from 1 (not red-listed) to 5 (critical). For taxonomic weights we used the root weight of May (1990) and Vane-Wright et al. (1991). For parameter p ...we used 0.3 throughout the study...The total weights... w_j , varied from 1 to 12.07, with an average of 2.16.

Apronen et al. (2005) do not discuss the scale type of regional W_j^R and national W_j^N rarity weights. However, given the information available it seems reasonable to assume that regional and national rarity weights are independent ordinal scales. (Taxonomic weights may also be ordinal. There is not space here to discuss the scale type of these cladistic-topology-based metrics. However, the following discussion does not depend on this determination.)

For the moment, let us assume that W_j^R and W_j^N are defined on a *common* 1 to 5 ordinal scale, and consider the following hypothetical data sample: For two species $j = 1, 2$: $W_1^R = 5$, $W_1^N = 1$, $W_2^R = 3$, $W_2^N = 2$, and $W_1^T = W_2^T = 1$. Assume $p = 0.3$. Then $w_1 = 2.2$ and $w_2 = 2.3$, and we conclude that species 2 has greater species weight than species 1 ($w_2 > w_1$). However, if rarity is an ordinal scale, then the following admissible (monotone) transformation produces another acceptable rarity scale: $\phi(1) = 1.2$, $\phi(2) = 2$, $\phi(3) = 3$, $\phi(4) = 4$, and $\phi(5) = 5$. With this scale $w_1 = 2.34$, w_2 is still 2.3, and $w_1 > w_2$. Order comparisons of species weights are meaningless.

Under certain reasonable assumptions, it has been shown (Ochinnikov 1996) that the only meaningful merging-functions for data measured with respect to a common ordinal scale are

order statistics such as the median. Apronen et al.'s (2005) w_j is not one of these. The situation for merging *independent*, ordinal scales is bleaker still. Again, under certain reasonable conditions, the only possible merging-functions are either the constant function or dictatorship (Aczél & Roberts 1989). In the case of species weights, dictatorship means that either regional or national rarity alone would determine the merged value. It is also worth pointing out that legitimately merged values for ordinal data are again ordinal. Thus calculating the average total species weight for median- or dictator-based w_j would be vacuous. More importantly, as presently defined, Apronen et al.'s objective function $F(X)$ is an inappropriate merging-function (based on results in Aczél & Roberts, 1989). The consequences of this fact for optimal reserve selection are not clear.

Apronen et al. (2005) recommend that “benefit functions and species weighting should be considered as standard options in reserve-selection applications.” The above discussion indicates there is still work to be done to integrate biodiversity value into meaningful conservation decision-support applications.

Conclusion

Meaningful statements “say something significant about the fundamental relationships among the objects being measured, whereas statements that are dependent on a particular, arbitrary choice of scale do not” (Roberts 1979).

Assessing the status of the world's ecosystems and measuring their change is a critical function of conservation science. Sparse data sets and system complexity have compelled conservation scientists to estimate data through expert judgement and other scoring, ranking, and rating procedures. To derive useful conclusions from estimated data may require facing the measurement-theoretic problems of determining scale type and assessing meaningfulness. Simple as well as sophisticated methodologies for prioritizing conservation efforts and acquisitions, applied in ignorance of these issues, may produce conservation choices justified by numerical artifact rather than empirical evidence.

Advocacy is an integral part of conservation science. Effective advocacy requires public confidence in the procedures underpinning policy recommendations. As Song and M'Gonigle (2001) point out,

striving for truth has long been the scientist's stock in trade and a source of authority for science in public policy. Scientists must provide an '*objective* contribution via the scientific method'...to facilitate sound decision making and rational administration.

Objectivity and scientific method are inextricably tied to measurement theory and practice. Scoring, ranking, and rating procedures do not automatically produce quantitative scales and neither does expert judgement. To presume otherwise is to risk the scientist's stock in trade. Only meaningful conclusions can "facilitate sound decision making and rational administration."

In a discipline where ground truth is often unknown or unknowable, measurement-theoretic meaningfulness guards against numerical arbitrariness. Understanding the limitations on conclusions from measurement scales can insure that conservation science is based on empirical relationships and can contribute to substantive understanding of the world.

Acknowledgements

I thank J. Williams for his extensive comments on an earlier draft of this manuscript. This research was supported by Wilburforce Foundation and Packard Foundation grants.

Literature Cited

- Aczél, J. and F. Roberts. 1989. On the possible merging functions. *Mathematical Social Sciences* **17**:205-243.
- Arponen, A, R. Heikkinen, C. Thomas, and A. Moilanen. 2005. The value of biodiversity in reserve selection: representation, species weighting, and benefit functions. *Conservation Biology* **19**:2009-2014.
- Chrisman, N. R. 1998. Rethinking levels of measurement for cartography. *Cartography and*

- Geographic Information Systems **25**:231-242.
- Chrisman, N. R. 2002. Exploring Geographic Information Systems, 2nd Ed. John Wiley & Sons, New York, New York.
- Cook, W. D. and M. Kress. 1992. Ordinal Information and Preference Structures. Prentice-Hall, Englewood Cliffs, New Jersey.
- Davis, F. W., D. Stoms, and S. Andelman. 1999. Systematic reserve selection in the USA: an example from the Columbia Plateau ecoregion. *Parks* **9**:31-41.
- Dunn, E. H., D. Mussell, and D. Welsh. 1999. Priority-setting tool applied to Canada's landbirds on concern and responsibility for species. *Conservation Biology* **13**:1404-1415.
- Groves, A. 2003. Drafting a Conservation Blueprint Island Press, Washington D. C.
- Johnson, C. J. and M. Gillingham. 2004. Mapping uncertainty: sensitivity of wildlife habitat ratings to expert opinion. *Journal of Applied Ecology* **41**:1032-1041.
- Knapp, T. R. 1990. Treating ordinal scales as interval scales: an attempt to resolve the Controversy. *Nursing Research* **39**:121-123.
- Krantz, D. H., R. Luce, P. Suppes, and A. Tversky. 1971. Foundations of Measurement Theory, Vol. 1. Academic Press, New York, New York.
- Luce R. D., D. Krantz, P. Suppes, and A. Tversky. 1990. Foundations of Measurement Theory, Vol. 3. Academic Press, New York, New York.
- Margules, C. R. and R. Pressey. 2000. Systematic conservation planning. *Nature* **405**:243-253.
- Meffe, G. J. 1998. Editorial: Conservation scientists and policy process. *Conservation Biology* **12**:741-2.
- Meir, E., S. Andelman, and H. Possingham. 2004. Does conservation planning matter in a dynamic and uncertain world? *Ecology Letters* **7**:625-622.
- Ovchinnikov, S. 1996. Means on ordered sets. *Mathematical Social Sciences* **32**:39-56.
- Pekeč A. 1997. Optimization under ordinal scales: when is the greedy solution optimal? *Mathematical Methods of Operations Research* **46**:229-239.
- Pekeč A. 1996. Scalings in Linear Programming: Necessary and Sufficient Conditions for

- Invariance. Technical report. BRICS Research Series RS-96-48.
- Possingham, H., I. Ball, and S. Andelman. 2000. Mathematical methods for identifying representative reserve networks. Pages 291-305 in S. Ferson and M. Burgman, editors. Quantitative methods for conservation biology. Springer-Verlag, New York, New York.
- Pressey, R. L. 1994. Ad hoc reservations: forward or backward steps in developing representative reserve systems. *Conservation Biology* **8**:662-668.
- Pressey, R. L. and K. Taffs. 2001. Scheduling conservation action in production landscapes: priority areas in western New South Wales defined by irreplaceability and vulnerability to vegetation loss. *Biological Conservation* **100**:355-376.
- Roberts, F. S. 1979. *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*. Addison-Wesley, London.
- Roberts, F. S. 1994. Limitations of conclusions using scales of measurement. Pages 621-671 in S. M. Pollock, M Rothkopf, and A. Barnett, editors. *Handbooks in OR/MS, Vol. 6*. North Holland, New York, New York.
- Smith, P. G. and J. Theberge. 1987. Evaluating natural areas using multiple criteria: theory and practice. *Environmental Management* **11**:447-460.
- Song, S. J. and R. M'Gonigle. 2001. Science, power, and system dynamics: the political economy of conservation biology. *Conservation Biology* **15**:980-989.
- Stevens, S. S. 1946. On the theory of scales of measurement. *Science* **103**:677-680.
- Suppes, P., D. Krantz, R. Luce, and A. Tversky. 1989. *Foundations of Measurement Theory, Vol. 2*. Academic Press, New York, New York.
- Williams, J. C., C. ReVelle, and S. Levin. 2004. Using mathematical optimization models to design nature reserves. *Frontiers in Ecology and the Environment* **2**:98-105.
- Wilson, K. A., M. Westphal, H. Possingham, and J. Elith. 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation* **122**:99-112.