Assessing and Communicating Data Quality in Policy-Relevant Research

ROBERT COSTANZA*
Coastal and Environmental Policy Program
Center for Environmental and Estuarine Studies
University of Maryland
Solomons, Maryland 20688-0038, USA

SILVIO O. FUNTOWICZ
Commission of the European Communities
Joint Research Center
Institute of Systems Engineering and Informatics
I-21020 Ispra (VA), Italy

JEROME R. RAVETZ
The Research Methods Consultancy Ltd.
13 Temple Gardens
London NW11 0LP, UK

ABSTRACT / The quality of scientific information in policy-relevant fields of research is difficult to assess, and quality control in these important areas is correspondingly difficult to maintain. Frequently there are insufficient high-quality measurements for the presentation of the statistical uncertainty in the numerical estimates that are crucial to policy decisions. We propose and develop a grading system for numerical estimates that can deal with the full range of data quality—from statistically valid estimates to informed guesses. By analyzing the underlying quality of numerical estimates, summarized as spread and grade, we are able to provide simple rules whereby input data can be coded for quality, and these codings carried through arithmetical calculations for assessing the quality of model results. For this we use the NUSAP (numeral, unit, spread, assessment, pedigree) notational system. It allows the more quantitative and the more qualitative aspects of data uncertainty to be managed separately. By way of example, we apply the system to an ecosystem valuation study that used several different models and data of widely varying quality to arrive at a single estimate of the economic value of wetlands. The NUSAP approach illustrates the major sources of uncertainty in this study and can guide new research aimed at the improvement of the quality of outputs and the efficiency of the procedures.

The Problem of Data Quality in Policy-Relevant Research

In scientific research, as in any other sphere of activity, the maintenance of the quality of products is critical for their effective use. In matured fields of traditional science, quality control is exercised informally by the competent practitioners (Ravetz 1971). In most scientific studies, the scientists actually doing the analysis have a good working understanding of the inherent quality of their measurements and results, but there is no accepted method to communicate this knowledge of data quality to potential users of the information. When the results of research are intended to be used as inputs to public policy decisions, the users of this information must either be knowledgeable in the details of the research methods or accept the results with no idea of their quality. Usually, they lack the knowledge for performing their own assessment of quality and must do without. As a result, the policy process on very many issues is impaired. This deficiency is recognized in the courts, as when expert witnesses are used for technical questions, but the traditional criteria for the quality of personal testimony do not apply to scientific information and so the courts generally pursue the vain hope of scientific certainty. Worse yet, in the absence of a quality-assessment system, these deficiencies are largely unrecognized and their consequences are difficult to estimate. Grades for quality are routinely assigned in innumerable spheres of activity in our society; yet in the case of information, one of the most sensitive products we have, there are no standard systems for grading and hence no means for a socially effective system of quality control. This article presents a system aimed at rectifying this situation.

The standard techniques of statistics were developed to handle a different aspect of uncertainty. They assume that uncertainty is due to real, precisely measured variation in the populations being sampled. They generally assume that we have a probability distribution to work with, without asking how well we know that distribution. This article concentrates on the issue of how well we know the distribution (its quality) and how we

KEY WORDS: Uncertainty; Data quality; NUSAP; Environmental policy; Policy-relevant research

*Author to whom correspondence should be addressed.

may communicate that knowledge. The standard statistical assumptions are justifiable in the case of the traditional experimental or field sciences, but the data available in policy-related research is frequently so scattered and coarse that refined mathematical manipulations do not possess much genuine meaning when applied to them. The more basic problem is that of partitioning of the total uncertainty into that due to real variation in the population (statistical uncertainty) and that due to errors in measurement (data quality).

The achievement of scientific work of high quality requires the deployment of sophisticated craft skills, as well as the motivating force of commitment and morale. Therefore, a full specification of quality would be as complex and subtle a task as the research itself. Fortunately, from the inexpert user's point of view, the relevant aspects of quality depend more on the product than on the process. Since the product of research is information that has certain knowledge as its ideal, we can assess quality by that yardstick. The incompleteness of certainty (or the inevitable uncertainty) of scientific information can be used to define a system of quantitative estimates and qualitative grades. By this means the various aspects of the uncertainty, and hence of the quality, of scientific information can be described.

No scientific activity is free from uncertainty; we may say that the key to a science being matured in its success in the recognition, communication, and control of the various sorts of uncertainty that affect its results and predictions. These include inexactness, as expressed by significant digits, unreliability as expressed in systematic error, and others. No amount of sophisticated apparatus and computer power can replace theoretical understanding of the problems of uncertainty or the practical skills of controlling and communicating it.

When quantitative information is used to provide inputs for the policy process, as in the case of indicators in the social and environmental fields, the scientist's problems of management and communication of uncertainty are severe. First, the original data are rarely as well controlled as in the laboratory. Well-structured theories, normally expected to be available in basic or applied science, are conspicuously absent in policy-related research. Furthermore, such research is interdisciplinary, involving fields of varying states of maturity and with very different sorts of practice in their theoretical, experimental, and social dimensions. Scientists must use inputs from fields they do not know intimately, so they cannot make the same sensitive judgments of quality that they do in their own subject. The result is that the quality control on the research process is diluted; the quality assurance of results is weaker; and they command less confidence among users.

The problems of uncertainty in policy-related research are further increased in its public dimension. Science is judged by the public (including decision makers) by its performance in such sensitive areas as hazardous wastes, radioactive fallout, food additives, and reproduction engineering. All these involve much uncertainty and also inescapable social and ethical aspects. Simplicity and precision in predictions, or even in the assignment of safe limits, are not feasible; yet policy makers tend to expect straightforward information as inputs to their own decision-making process. They want their statistical indicators to provide them with certainty.

In such circumstances, the maintenance of confidence in science among policy makers and the general public becomes increasingly difficult. The issue manifests itself at several levels. The simplest is in the representation of uncertainty in quantitative estimates. For example, in risk assessments, the scientist who advises policy makers knows that a prediction like a one-in-a-million chance of a serious accident should be qualified with statements about the many sorts of uncertainty, so as to caution any user about the limits of reliability of the numerical assertions. If these are all expressed in prose, the statement becomes tedious and incomprehensible to the lay users; if they are omitted or even given in some simple statistical representation, then the same advisor can be accused of conveying a certainty that is not warranted by the facts. Lacking a means for qualifying his quantitative statements, the advisor is caught in a communications trap.

Yet another even worse dilemma is encountered whenever scientists give advice on policy-related issues. In addition to low-frequency hazards, these may concern diffused hazards, such as pesticides or food additives, or possible large-scale future environmental perturbations, such as the greenhouse effect. Such advice is usually supported by the present or expected behavior of some critical indicator. This dilemma has been stressed in a recent article by Maddox (1987) "Half-truths make sense (almost)". This was a comment on a prediction of the consequences of the greenhouse effect, using the rise in global mean temperature by 2030 AD as the indicator. Offering definite advice is a risky business: a prediction of danger will appear alarmist if nothing happens in the short run, while reassurance can be condemned if it retrospectively turns out to be incorrect. Then the credibility of science, traditionally based on its supposed infallibility, is threatened. Facing this dilemma, the scientific advisor may prudently refuse to accept low-quality expert opinions as a basis for quantitative threshold values and consequently decline to provide a definite opinion when requested by policy makers. Then science is perceived as failing to perform its public function of offering advice and assurance, and its legitimacy is again threatened. Thus, the credibility of science, based for so long on the supposed certainty of its conclusions, is endangered by *any* sort of statement on such inherently uncertain issues, be it advice or disclaimers. Is there no way to escape the horns of this dilemma, in which the credibility and the legitimacy of science are both at risk? As Maddox (1987) puts it, these are among the trials with which policy research centers must contend. "Tell the people that there is a muddle, or give them a clear message that they must man the barricades?" The World Resources Institute's solution to the dilemma (the adoption of a computer model) is described by Maddox as a "cop-out."

Thus, in policy-related research, the traditional tools for assuring quality through the control of uncertainty need to be enriched. To solve the scientific advisors' dilemma, one needs something more than bigger, faster computers and more data. Uncertainty is integral to these problems; it cannot be removed by technical means. It must be managed and effectively communicated so that it becomes a recognized input to the decision-making process. We should exhibit the structure of our uncertainty, so that the quality of our information in relation to its functions is assured. The notational system NUSAP (Funtowicz and Ravetz 1991) has been designed and developed to further the evolution toward these ends.

Science and the Management of Uncertainty

As natural science has grown and matured over the centuries, it has developed tools for the management of different sorts of uncertainty. Each particular set of tools was devised in response to a recognized problem. Quantitative measurements have been made since antiquity in such fields as astronomy, but not until the early nineteenth century was an effective *calculus of errors* created. In a separate tradition, *combinatorial probabilities* were created in the seventeenth century for the analysis of games of chance. These mathematical tools were then available in the later nineteenth century for use in all natural sciences involving random processes. A parallel development was in *statistics* from the seventeenth century onwards, involving aggregated information gathered for its importance to statecraft and commerce. These three approaches can be seen to relate to different aspects of the limits of our knowledge: errors relate to the limits of exactness of measurements made with real instruments; randomness relates to the limits of causality and determinism as observable in the natural world; and statistics relates (implicitly in its practice) to the limits of correspondence between descriptive categories and the reality to which they relate. These three approaches have all interacted with and enriched each other, so that now they are not distinct in name or subject matter.

Another way of looking at the history of the management of uncertainty is in terms of the relation between the researcher and the system under study. Combinatorial probabilities describe an abstract world, where knowledge of processes is incomplete, but where events occur totally independently of the observer and have no imprecision in themselves. The theory of errors arose from the realization of the interaction between the elements of the measurement process. These are the instruments with a limited fineness of scale and accuracy of construction and calibration, and the individual operators with their perceptual inaccuracies and distortions. In statistics there is an analogous effect; general concepts (e.g. population, income) must have operational definitions, and data collection and analysis must have safeguards against a variety of possible errors. A scientific underpinning for error theory has been provided by quantum mechanics, with its proof that each observation has a finite effect on the system and that there is a lower bound to the imprecision of related pairs of measured magnitudes.

Within the present century another approach to probability has been articulated, which formalizes personal judgements of the confidence to be placed on assertions and (in principle) allows their calculation by means of Bayes' theorem on inverse probabilities (Savage 1954). Such subjective probabilities have been applied to the analysis of complex problems in risks and the environment, where (as elicited from experts) they are used to remedy the lack of reliable statistical data and established causal theories. The formal system of those subjective probabilities is equivalent to that of fuzzy sets (Gupta and others 1979). These may be seen as managing linguistic uncertainty. Thus, old does not have a distinct boundary with not old or young. The function describing membership of human ages in the class old will have the value zero for 1 and 2, and unity for 99 and 100. An intermediate zone of ages, describable as not so old, pretty old, etc., will have intermediate values for the fuzzy-set membership function (Zadeh 1965). Since it is formally equivalent to the traditional probability calculus, fuzzy-set theory enables an elaborated computation with subjective probabilities.

These techniques have been developed mainly in connection with the study of highly articulated models of decision procedures, in the field of policy analysis (Morgan and Henrion, 1990), rather than for the characterization of the highly uncertain information that is common in policy-related research. For the latter case (which is our present concern), they are well adapted for a finer description of the inexactness or spread of estimated quantities, since they can exhibit a complete probability distribution rather than a simple mean and variance, however, such information is not at all certain. To characterize its own uncertainties (quite a real problem in the case of what are only personal opinions) leads to marked difficulties in the theory because the only means available within the framework of the theory is to produce a subjective probability distribution for a subjective probability distribution, and it is not easy to find a straightforward interpretation for such a second-order operation (Henrion 1988). In our scheme, this function is performed by a separate category, the *assessment* or *grade*, which directly expresses the reliability of the information rather than standing for its iterated probability.

All the formal approaches to probability are afflicted by the problems of applying analytical techniques for the management of the severe uncertainties in the information that is frequently the best available in policy-related research. All the traditional statistical techniques, even the non parametric methods for ordinal data, presuppose that the data are numerous and of good quality. In the face of a few coarse and scattered readings, all such techniques require many strengthening assumptions about the data for their applications to be legitimate. The significance of low-level environmental effects, as assessed by standard tests, may well depend more on computing procedures and on hypotheses about the shape of the data distributions than on the scanty data themselves. (Bailar 1988, pp. 8–12).

It is because of the inadequacy of the best attempts to apply traditional statistical techniques to the highly imperfect data so frequently encountered in policy-relevant research that we have developed the coarser, semi-qualitative approach that is advanced here. Our approach:

1. avoids the attempt to replace human judgements by formal systems or by computer programs—instead it applies simple yet robust techniques for the guidance of those judgements;
2. relies more on description and classification than on formal calculations;
3. complements the standard statistical techniques as commonly taught and practiced.

It can be useful even in cases where the data are appropriate for manipulation by the standard techniques; but we believe that it is essential in the circumstances of policy-related research.

As Bailar (1988) puts it:

> All the statistical algebra and all the statistical computations are of value only to the extent that they add to the process of inference. Often they do not aid in making sound inferences; indeed they may work the other way, and in my experience that is because the kinds of random variability we see in the big problems of the day tend to be small relative to other uncertainties. This is true, for example, for data on poverty or unemployment; international trade; agricultural production; and basic measures of human health and survival.
>
> Closer to home, random variability—the stuff of P-values and confidence limits, is simply swamped by other kinds of uncertainties in assessing the health risks of chemical exposures, or tracking the movement of an environmental contaminant, or predicting the effects of human activities on global temperature or the ozone layer. It was, in fact, this aspect of environment problems that first attracted me to the field. I have long had an interest in non-random variability, and here I see it in almost pure form [pp. 19–22].

In most discussions among scientific researchers actually performing measurements there is usually at least an implicit recognition of the relative degree of quality of various methods, but once a number is obtained and exported to the general scientific community or to an inexpert policy audience, it is implicitly presumed good. Even if it is quoted with its known imprecision (e.g., by significant digits or by mean and variance), there is no record of its quality. Only those experts familiar with the details of the measurement methods will remember it.

In the case of policy-related research, most of those using the information are not experts in its production; they will not know, or be able to understand, the technicalities of the research that allow its quality to be understood. They will at first take the information on trust, presuming it to be of good quality; if this turns out to be specious, then the credibility of all comparable information can be impaired. All experts tend to be accorded equal credibility in policy debates. When any one is discredited, the credence given to all and the credibility of science itself are diminished. The communication of data quality therefore must be clear and easily understood if it is to be at all effective.

We will use a familiar example to show how a single numerical measure is not adequate for expressing all aspects of the quality of information. A marksman shooting at a target will produce a pattern of shots. They may all cluster tightly, in which case we speak of high precision, but we are also concerned with how closely they come to the bullseye, which we describe in terms of accuracy. If the sighting apparatus is defective, the marksman's shots may well have high precision and low accuracy. (Of course, precision instruments are those with both high precision and high accuracy). In general scientific practice, the thing being measured is not accessible to such simple, naked-eye observation; and while precision can be simply measured (as by the variance of a set of readings, what we will refer to as the spread), accuracy is known only indirectly, by the accumulated experience of the behavior of the system. Under such circumstances, the equivalent to accuracy becomes the reliability of the process, or the strength of the results, or what we will refer to as the grade. Its assessment is accomplished by judgements based on craft experience rather than by counts and calculations.

It is important to appreciate that the two attributes are quite independent. Using the example of the marksman with defective sighting apparatus, one can see how there could be high precision (low spread) with low accuracy (low grade). On the other hand, there are many scientific situations where the best available data may be of relatively low precision (high spread), but can nonetheless be quite reliable as a representation of the thing being measured and therefore deserve a high grade. High quality in scientific information is not secured by the banishment of uncertainty (for that is impossible) but by its effective communication and management.

NUSAP Notational System

Our notational system is thus based on an elucidation of these two sorts of uncertainty. We start with the simplest sort, usually expressed by error bars and significant digits. Every set of data has a spread; it is an attribute of any quality, however derived; it may be considered in some contexts as a degree of precision, as a tolerance, or as a random error in a calculated measurement. It is the kind of uncertainty that relates most directly to the quantity as stated and is most familiar to students and even the lay public.

A more complex sort of uncertainty relates to the level of confidence to be placed in a quantitative statement; this relates to the accuracy that we have contrasted to precision. In statistical practice, this is usually represented by the confidence limits (at, say, 95% or 99%). In practice, such judgments are quite diverse; thus safety and reliability estimates are given as conservative by a factor of n. In risk analyses and futures scenarios, estimates are qualified as optimistic or pessimistic. In laboratory practice, the systematic error in physical quantities, as distinct from the random error or spread, is estimated on an historic basis. Thus, it provides a kind of assessment to act as a qualifier on the number, or alternatively (if desired) on the spread. We call this attribute the *grade* to convey the qualitative degree of goodness of a number. This assessment of grade is one level up from spread, both in its sophistication and variety. Our knowledge of the behavior of the data gives us the spread; and our knowledge of its production or intended use, gives us its grade.

We can now introduce the full notational system designed for the management of uncertainty in quantitative information. We call it NUSAP; the last three letters in the acronym refer to the spread, assessment and pedigree. The first two refer to numeral and unit. The first category encompasses the arithmetical system, and the second the base in which it is appropriately expressed. In a full NUSAP expression, there is a balance between all the elements; thus the number of significant digits in the numeral place, when combined with the scaling factor in the units place, will be coherent with the inexactness described under spread.

We arrive at the grade by means of an evaluative accounting we call the *pedigree* of the data. We have developed a matrix of items that show the boundary with ignorance (Table 1) by displaying the degrees of strength of crucial theoretical, empirical, and social components of the process. The theoretical, empirical, and social components are *quality of models, quality of data*, and *degree of acceptance* (Table 1). By scoring a particular measurement in each of these components we can describe its pedigree, and this pedigree is used to assess the measurement's grade. The three components are coded on an ordinal scale of 0–4, and their average, normalized on the scale 0–1, provides a convenient measure of the overall grade. It should be clear that this scale provides a simple and suggestive index, and not a measured quantity. Provided that it is used with that awareness, and is not embedded in complex, hyperprecise mathematical manipulations, it will function as a useful tool in the evaluation of scientific information. Thus the pedigree, in exhibiting the limits of the state-of-the-art of the field in which the information was produced, provides us with a simple gauge for an assessment of the strength of that information, or its grade.

For example (referring to Table 1), if we qualify the theoretical component of a particular measurement as having used a computational model, we are implicitly stating that we do not have a theoretical model. We thus record the absence of an effective theory and score the theoretical component as a 2. Similarly, if the empirical component is not experimental, it can be at best historical or field data, as in most environmental research, which would be scored as a 3. In the latter case, data are inherently less capable of control; so it is less effective as an input and check on the quality of the model. The components on the social side describe the evaluation of

Table 1. Numerical estimate pedigree matrix

Score	Theoretical, quality of model	Empirical, quality of data	Social, degree of acceptance
4	Established theory Many validation tests Causal mechanisms understood	Experimental data Statistically valid samples Controlled experiments	Total All but cranks
3	Theoretical model Few validation tests Causal mechanisms hypothesized	Historical/field data Some direct measurements Uncontrolled experiments	High All but rebels
2	Computational model Engineering approximations Causal mechanisms approximated	Calculated data Indirect measurements Handbook estimates	Medium Competing schools
1	Statistical processing Simple correlations No causal mechanisms	Educated guesses Very indirect approximations "Rule of thumb" estimates	Low Embryonic field
0	Definitions/assertions	Pure guesses	None

the information in its particular context. Degree of acceptance of a result will be straightforward in a fully matured field where criteria of quality are agreed; a rough approximation to this is the referee's judgment on the research paper. The social degree of acceptance component is required because it represents an additional element in the overall pedigree of uncertainty. Measurements with high theoretical and empirical quality can still have high uncertainty if they have not withstood the tests of peer review and achieved scientific consensus.

NUSAP is a system because it is not simply a collection of fixed notations. Rather, it is a set of determinate categories, each of which can be filled by particular notations appropriate to the particular context of application. The names of the five categories (or boxes, or places in a string) make up the acronym NUSAP. By means of this place value representation, each category can be expressed without need for its explicit identification (this is why we refer to it as a scheme of notations). For each category, there are many possibilities for conveying particular desired meanings; thus in energy studies, kilowatts, megawatts and gigawatts are not merely the same unit with prefixes denoting a scaling increasing by a thousand. Rather, they refer to physical and accounting operations at very different levels, and they have quite different meanings as indices. Thus, writing in the NUSAP notation we do not have 1 kW = 10^{-6} GW, except in a formal arithmetical sense. Similarly, units of money have various appropriate expressions that are clearly distinguished in NUSAP. Thus, for the valuation of wetlands a common unit is: thousand dollars per acre at a particular date. So we write "$K/acre/yr" in the unit place. These cases show how NUSAP representations can convey variants of meanings that may appear not too different in ordinary practice but that are conceptually quite distinct. Any particular array of such constituents in the five places we call a notation. Given such a notation, any particular case, such as "6 1/2 $K/acre/yr" will be an "instance" of the notation.

The key distinction between NUSAP and other notational systems is its incorporation of more of the components of uncertainty, in particular the assessment of the grade of the estimate. It is not intended as a final and complete system for this purpose, but as a starting point in the evolution of our management of uncertainties in data quality to compliment our management of other forms of uncertainty.

To show the use of NUSAP for characterizing uncertainty in information in policy-related research, we take the example criticized by Maddox (1987). The chosen indicator (temperature rise consequent on greenhouse effect) as described by the World Resources Institute did not exhibit effective control of its uncertainties, and so [as Maddox (1987) said] could not command very high confidence. The original statement was of a rise between 1.6°C and 4.5°C (in average earth temperature over the next 40 yr). In NUSAP, this increase is best displayed as,

3	°C	±50%	[0.5]
numeral	unit	spread	grade

The first three places are derived directly from the quoted quantitative prediction. The range of $1.6x$ to $4.5x$ is near enough to 3 ± 50%, and it may be said that our mode of representation is more faithful to the scientific meaning of the datum. It thus displays skill in the management of uncertainty and would, therefore, tend to maintain confidence in the scientists making the prediction. We estimate a pedigree of (2, 2, 2) using the model pedigree matrix of Table 1, and this translates to a grade of [0.5] [(2 + 2 + 2)/12]. This rating comes from analyzing the World Resources Institute's model of global warming. Are there effective theoretical models for atmospheric CO_2 and its temperature effects? According to Maddox (1987), not quite; severe uncertainties exist on the time scales both of millions of years and of days. Hence we have, at best, computational models. What about the data that are injected into the models as inputs? There are some that are better than educated guesses but (as yet) not much obtained through instrumental readings, even as historic/field data. Hence we do best to call them calculated data, usually resting indirectly on measurements. Moving now to the social aspects of the pedigree, the degree of acceptance of the result seems to be medium. The criticisms of Maddox (1987) are of the policy implications of the prediction rather than of the quantity itself.

This mapping of the limits of the state-of-the-art exhibits the boundaries with our ignorance. We do not know as much as we could had there been theoretical models and historic/field data. In such a case, predictions would have had more strength, more justified urgency, and perhaps also more information on environmental consequences and remedies. As it is, our knowledge is not quite swamped by our ignorance, but it is still too weak and unfocused for effective decision making. Our ignorance in the policy aspects of the problem has scarcely been dispelled. Thus, the indicator of the World Resources Institute, when properly expressed, tells us more about our ignorance than about the biosphere.

In this way we see how relevant evaluations of quality are expressed through the assessment and pedigree categories. These are cast in terms of the characteristic uncertainties of the information, including the border with ignorance; and they can be expressed in a form most appropriate for the policy problem that defines the indicator itself.

As we have seen, the NUSAP system enables the representation of a quantity (with its uncertainties) in a variety of ways. Each form of expression corresponds to a particular type of message, depending on the content and context to be conveyed. Some expressions may contain more detailed information than others, if the loss of brevity is justified by the gain in effectiveness. The full NUSAP form, as given above, is the most general framework for such expressions. In it, the assessment box may be used to constitute the grade or degree of goodness. However, when a set of similar numbers are being compared, some of the boxes may be redundant. For many policy makers it is sufficient to use the most abbreviated form, the pair (N, A), where N (the numeral) is a representative number, and A (the assessment) is a code for the grade that describes the degree of goodness of the number, as distinct from its spread.

An Arithmetic for Data Quality

The NUSAP system is designed to enable judgements of the quality of quantitative information to be applied in a consistent manner, so that all users of that information can make the same (or roughly the same) evaluations of it. So far we have shown how the various aspects of the uncertainty of information can be described in standard codes; the system is thus established as a descriptive scheme. As such, it has a limited usefulness. The numbers by which quantitative information is described are also used for calculation, and the NUSAP system should enable the various sorts of uncertainty to be tracked through a computation. Then, from given uncertainties in the inputs, it should be possible to describe the uncertainties in the outputs of a calculation. With that accomplished, the NUSAP system would constitute a true arithmetic, a special one for the characterization, manipulation, and ultimately the control of uncertainty in quantitative information. For this first sketch of an arithmetic of uncertainty, we will restrict ourselves to the case of a small set of elementary operations. There may well be problems arising when operations are iterated many times, for which further special rules may be necessary.

An arithmetic of uncertainty will be slightly different from ordinary arithmetic in several ways. First, since it is dealing with particular properties of numbers, its operations will not necessarily mirror those affecting the numbers themselves. Moreover, we can expect that those concerning the spread will not be the same as those concerning the grade, since one is primarily a quantitative property while the other involves judgements of quality. In addition, there will be some modifications of the rules to fit special cases. These can be detected by examination of the numbers involved (this aspect of the arithmetic can be programmed on a computer, along with the basic arithmetic).

It might be thought that an arithmetic with exceptional cases is a rather odd arithmetic; hence it may be useful to show the sorts of special cases affecting ordinary arithmetic, which in themselves indicate why a special notation for uncertain quantities is necessary.

Consider two simple sums:

$$\begin{array}{rr} 1{,}000{,}000 \\ +\underline{5} \\ 1{,}000{,}005 \end{array} \qquad \begin{array}{rr} 1{,}000{,}000 \\ +\underline{5} \\ 1{,}000{,}000 \end{array}$$

The first of these is in conventional arithmetic; it is arithmetically correct, but probably nonsense in practice.

For, except in the very rare case when the million has been counted out with perfect precision, the quantity 5 is vanishingly small in comparison with it. On the other hand, the second sum, while meaningful in practice, is formally incorrect as arithmetic; one can imagine the reaction of a schoolteacher if a pupil produced and defended such a sum! This pair of sums is the simplest representative of a large class of arithmetical operations where the numbers do not, or cannot, mean what they seem to; in technical language, we say that the zero is an ambiguous symbol, capable of standing for either a digit (although even then a strange one) or a filler of an empty place denoting a larger unit of aggregation for counting. In NUSAP notation, the sum in its simplest form would read as follows:

$$1 \text{ E6} + 5 \text{ E0} = 1 \text{ E6}$$

and it is clear that the two quantities are of different orders of magnitude, incapable of addition in an ordinary sum.

Let us next consider the rules of simple arithmetic that are appropriate for spread. These follow the traditional rules of the simple calculus of errors: in sums and differences, absolute errors add; while in multiplications and divisions, proportional errors add. We could adopt the common rule that the root mean square sum of errors is to be taken, but this would introduce an inappropriate degree of complexity of calculation at this stage. Thus (with a, b, c, d all positive)

$$(A \pm a) + (B \pm b) = (A + B) \pm (a + b)$$
$$(A \pm a) - (B \pm b) = (A - B) \pm (a + b)$$

and

$$(C \pm c\%) * (D \pm d\%) = (C * D) \pm (c + d)\%$$
$$(C \pm c\%)/(D \pm d\%) = (C/D) \pm (c + d)\%$$

The use of percentages to express proportional spreads must be done with caution, as it is very easy to write meaningless percentages such as 102% or 700%. When spreads are as large as, or larger than, the number itself, then the expression of proportional spreads requires some skill. Quite reasonable percentages can give quite large spreads; thus spreads of ±33%, ±50% and ±67% produce variations through factors of 2, 3, and 5, respectively (since, for example, $[1 + 67\%]/[1 - 67\%] = 5$). For larger proportional spreads than those, we should use a notation for "factor of", as F 10, to indicate variation through that proportional range.

The rules of elementary arithmetic for the grade are nearly as simple, although there are two exceptional cases to be observed. In the case of addition and subtraction, we usually take the weighted mean of the separate grades of the numbers. This reflects the intuitive judgement that the quality of the result should be the "average" grade of the collection. The strengths or weaknesses of the separate elements are given their influence, proportional to the size of that element. Using brackets to denote the grade, we have:

$$E,[e] \pm F,[f] = (E \pm F), [(E * e + F * f)/(E + F)]$$

The two exceptional cases both apply when the two terms are nearly equal. The reason for an exceptional grade is easier to see in the case of subtraction. If we have two terms that are nearly equal, say 95 and 92, then any uncertainties in the initial terms will be magnified in their difference. This is easy to see in the case of spread; if each has a spread of ±1, then their difference will be 3 ± 2. The proportional spread goes from about ±1% for each of the initial terms, to ±66% for the difference, quite an enormous change. It is hard to imagine such a number being in any way as reliable as either of the initial terms. Hence we must construct a rule, inevitably somewhat arbitrary in its details, for reducing the grade of the difference element when its spread is so dramatically increased. This will be a more general, simple, and coarse version of the rules for distinguishing the means of statistical distributions. We divide the rule into three cases. There is no change when the ratio difference to (average) spread (both expressed as percentages or as numbers) is greater than 5. If that ratio is less than 2, the grade is reduced by 50%. In between, the reduction is linear, bearing in mind that grades are expressed to the nearest single digit only.

The exceptional rule for addition is not quite so completely automatic in its operation; it comes into play when two quantities that are derived from independent procedures are averaged or compared in some other way. In this case, there is a qualitative judgement that if two such quantities are equal, or nearly so, by some appropriate criterion, then this property serves as a corroboration of them both. Even if neither of them can be checked directly against the reality that it is intended to measure, the unlikelihood that both have come to the same erroneous estimate serves as positive evidence that they are both more likely to be correct. For this case, we can apply the above rule in the other direction: when the ratio difference to (average) spread varies between 5 and 2, the grade is increased by up to 50%.

For multiplication of numbers, the rule for grade is simple; here we adopt a weak-link principle: the grade of the product is the minimum of the grades of the factors. Thus

$$G,[g] * H,[h] = G * H,[Min(g,h)]$$

For this rule, there are no exceptional cases.

Table 2. Summary of wetland value estimates (1983 dollars) for various components of wetlands contributing to their economic value, using two competing models [willingness to pay (WTP) and energy analysis (EA)] from Costanza and others (1989)

Method	Annual value per acre	Per acre present value at specified discount rate 8%	3%
WTP-based			
Shrimp	10.85	136	362
Menhaden	5.80	73	193
Oyster	8.04	100	268
Blue crab	0.67	8	22
Total commercial fishery	$25.37	$317	$846
Trapping	12.04	151	401
Recreation	3.07	46	181
Storm protection	128.30	1915	7549
Subtotal	168.78	$2429	$8977
Option and existence values	?	?	?
EA-based			
GPP conversion	$509–847	$6,400–10,600	$17,000–28,200
Best estimate	$169–509	$2,429–6,400	$8,977–17,000

As we have been displaying the various rules for combining spreads and grades, we have been showing the way in which numbers may be expressed using the full NUSAP system. In this simple format, the difference from ordinary notation is slight; thus we have, as a sample, 5 g/s ±2, [0.6]. Should it be useful to include the pedigree in the expression, then this might be done with another sort of bracket, giving an expression like 5 g/s ±2, [0.6], {2,3,2}.

We notice that the grade almost always decreases in calculations, sometimes quite drastically. It could be that many computations that are now accepted as reasonable and as giving meaningful outputs would, under such a grading system, be seen to be of very low quality. In particular, matrix-inversion operations, involving the sums and differences of many-factored products, would be especially vulnerable. The fault, however, might not lie in the peculiarities of a grading system, but rather in a class of mathematical operations over which there has hitherto been very little effective quality control. If these grading rules turn out to be too harsh in practice, then they can easily be modified in their details, but the principle, on which effective quality control can be based, that there must be some standard procedures for quality assessment, is not to be compromised.

An Example: Valuation of Ecosystems

To demonstrate the usefulness of the proposed system, we carry it through for the example case of ecosystem valuation. We use a well-documented study of the economic value of wetlands in Louisiana (Farber and Costanza 1987, Costanza and others 1989) that employed a number of different models and methods to arrive at an estimate of the total value of the ecosystem. The results from the original study are reproduced in Table 2.

There are two overall methods whose results are presented. The willingness to pay-based method enumerates the various components of ecosystem value and derives an independent estimate for each one. These components are then added to yield the total value. For example, shrimp production value was estimated as $10.85/acre/yr, and storm protection value as $128.30/acre/yr. Option and existence value are known to be important components of the total but no direct estimate was made for this ecosystem.

A second method (energy analysis) uses the total solar energy captured by the ecosystem as an indicator of its economic value. It is more comprehensive (in that it does not require summing individually measured components to arrive at the total), but the connection between energy captured and economic value is controversial and uncertain.

Finally, the present value of the ecosystem services are calculated using various discount rates based on the assumption that the ecosystems provide a constant stream of benefits into the indefinite future. In this case: present value = annual value/discount rate. The appropriate discount rate to use in such a situation is highly uncertain, however.

Table 3 is a recasting of these results into the NUSAP system. Here the numerical results are given only to the

Table 3. NUSAP scores and summary grades for elements of wetland valuation problem

Element	Numeral, N	Unit, U	Spread, S	Assessment	
				Pedegree	Grade
WTP-based estimates					
Shrimp	1 E1	$/ac/yr	±10%	(3,3,3)	0.7
Menhaden	6 E0	$/ac/yr	±20%	(2,2,2)	0.5
Oyster	8 E0	$/ac/yr	±30%	(2,3,2)	0.6
Blue crab	1 E0	$/ac/yr	±40%	(3,2,3)	0.6
Total Commercial Fishery	**2.5 E1**	$/ac/yr	**±20%**		**0.6**
Trapping	1.2 E1	$/ac/yr	±30%	(2,2,2)	0.5
Recreation	3 E0	$/ac/yr	±10%	(3,4,3)	0.8
Storm Protection	1.3 E2	$/ac/yr	±20%	(2,3,2)	0.6
Subtotal	**1.7 E2**	$/ac/yr	**±20%**		**0.6**
Option and Existence Values	5 E2	$/ac/yr	±50%	(1,0,1)	0.2
Total WTP	**7 E2**	$/ac/yr	**±40%**		**0.3**
EA-based					
GPP conversion	7 E2	$/ac/yr	±25%	(3,2,1)	0.5
Average of two methods	**7 E2**	$/ac/yr	**±30%**		**0.6**
Discount rate	5 E0	%	±50%	(1,3,1)	0.4
Present value	**15 E3**	$/ac	**±80%**		**0.4**

appropriate degree of precision, and the spreads on each number are shown (using only 10% increments except for 25% and 75%). The pedigree for each number is given, based on an analysis of the individual models and methods used (see Costanza and others 1989, for a complete description of these) coded using the 0–4 system in Table 1. For example, the shrimp production estimate was based on a theoretical model relating wetland area to shrimp catch (score = 3) using historical and/or field data from National Marine Fisheries shrimp catch statistics and measured wetland area (score = 3) in a procedure (regression analysis) that has high but not total peer acceptance for the intended purpose (score = 3). Finally the grade for each estimate is given based on the average scores in the pedigree [(3 + 3 + 3)/12 = 0.6]. Note that grades are rounded to one digit.

Several quantities are calculated in the table using the NUSAP arithmetic described above. These are shown in bold. The total commercial fishery value is the sum of four components. Its spread is the weighted average of the percentage spreads of the components [$(1E1 * 0.1 + 6E0 * 0.2 + 8E0 * 0.3 + 1E0 * 0.4)/2.5E1) = 0.2$]. Its grade is the weighted average of its component grades [$(1E1 * 0.7 + 6E0 * 0.5 + 8E0 * 0.6 + 1E0 * 0.6)/2.5E1 = 0.6$].

An estimate for option and existence value is given based on studies of other areas, but, as its spread and grade indicate, for this application it is definitely an order-of-magnitude estimate. The total WTP based value reflects the quantitative importance of option and existence values and their relatively low quality. We end with a spread of ±40% and a grade of 0.3 for this estimate.

The EA-based estimate yielded a very similar quantity estimate to the WTP estimate, and this is taken as corroborating evidence since the likelihood that this would occur by chance is small. The average of the two methods is therefore of higher grade than either of the inputs [0.6 vs (0.5 and 0.3)], and we are left with a reasonably high-quality estimate of the total annual value of wetland production [7 E2 $/acre/yr ± 30% (0.6)].

Converting this to present value significantly reduces the data quality, however, because of the high uncertainty about the discount rate. The spread on the present value goes to ±80% and the grade goes down to 0.4.

The two tables present very different pictures of the situation. This is partly because in the original table, the per acre present value was calculated for each component of ecosystem value, so that the original table has three columns of figures. In the NUSAP version, there is only one column, and the quantitative information is reduced to the minimum relevant to the quality assessment. Thus (for these purposes) we do not need to calculate present values with 3% and 8% discount rates, but only to provide a discount rate of 5 ±50%, with an appropriate pedigree. Hence the numeral column of the table corresponds generally to the first column of the original table.

We notice that much more information can be provided by the NUSAP approach. For instance, if someone consulting the table wanted to know about the quality of one of the estimates, then the number 10.85 (annual value for shrimp) would not convey very much useful information. Clearly, this is not intended to say that an estimate of 10.88 would be seriously wrong; but whether 12 would be equally acceptable is not clear. Thus the digits by themselves do not provide an estimate of their spread; the significant digits convention breaks down here as in many other cases, but with the NUSAP notation, the spread is given explicitly.

With the confidence that NUSAP affords in the handling of highly inexact quantities, we can include useful quantitative arguments that would otherwise be obscure or burdensome. For example, in a table of entries given to four or five digits (as the original one), there is no way to express the very inexact estimate of option and existence values. In the original table that row is left empty, with only question mark signs, but with NUSAP, this can be expressed as 5E2 $\pm 50\%$, or lying between 250 and 750, thus varying through a threefold range. When added to the willingness to pay estimates, this provides a sum of 7E2 $\pm 40\%$, and this compares well with the completely independent estimate of 7E2 $\pm 25\%$ by GPP conversion. It is then possible to invoke the principle of corroboration and to increase the grade of the common estimate.

Thus, the NUSAP representation of the series of calculations that went into the estimation of the value of wetlands offers a clear picture of the data quality. It also allows the uncertainty in the final estimate to be easily communicated, and it directs research to those areas most likely to improve the quality of that final estimate. A system such as this (should it come into general usage) would allow much saner management of our intellectual resources and much better management of our natural resources.

Acknowledgments

This article benefited from a number of informal reviews and discussions among our colleagues at Maryland, Ispra, and in the UK, including R. Ulanowicz, B. Hannon, J. Bartholomew, S. Tennenbaum, and B. de-Marchi. Partial support was provided by the National Science Foundation, grant BSR-8906269 titled: "Landscape modeling: The synthesis of ecological processes over large geographic regions and long time scales," R. Costanza and F. H. Sklar, Principal Investigators, and by the Joint Research Center of the Commission of the European Communities. We are indebted to T. O'Riordan, R. Christofferson, and one anonymous reviewer for their helpful suggestions on an earlier draft.

Literature Cited

Bailar, J. C., III. 1988. Scientific inferences and environmental problems: The uses of statistical thinking. Institute for Environmental Studies, University of North Carolina at Chapel Hill.

Costanza, R., S. C. Farber, and J. Maxwell. 1989. The valuation and management of wetland ecosystems. *Ecological Economics* 1:335–361.

Farber, S., and R. Costanza. 1987. The economic value of wetlands systems. *Journal of Environmental Management* 24:41–51.

Funtowicz, S. O., and J. R. Ravetz. 1991. Uncertainty and quality in science for policy. Kluwer, Dordrecht.

Gupta, M., R. Ragade, and R. Yager (eds.). 1979. Advances in fuzzy set theory and applications. North Holland, Amsterdam.

Henrion, M. 1988. Uncertain information processing in knowledge support system. *In* A. P. Sage (ed.), Concise encyclopedia of information processing in systems and organizations. Pergamon, Oxford.

Henrion, M., and B. Fischhoff. 1986. Assessing uncertainty in physical constants. *The American Journal of Physics* 00:000.

Maddox, J. 1987. Half-truths make sense (almost). *Nature* 326:637.

Morgan, M. G., and M. Henrion. 1990. Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press, New York.

Ravetz, J. R. 1971. Scientific knowledge and its social problems, Oxford University Press, New York.

Savage, L. J. 1954. The foundations of statistics. John Wiley, New York.

Zadeh, L. A. 1965. Fuzzy sets. *Information and Control* 8:338.