WORKING PAPER 499

IMPERFECT QUANTIFICATION IN GEOGRAPHY: FRAMEWORKS FOR QUALITY ASSESSMENT AND DISCLOSURE

S.M. MACGILL

School of Geography
University of Leeds
Leeds LS2 9JT

June 1987

IMPERFECT QUANTIFICATION IN GEOGRAPHY: FRAMEWORKS FOR QUALITY ASSESSMENT AND DISCLOSURE

S M Macgill, School of Geography, University of Leeds, Leeds LS2 9JT.
June 1987

ABSTRACT

Quantitative knowledge in geography, as with that in probably all other disciplines, is almost inevitably imperfect. In this paper we take stock of some of the key sources of uncertainty, ambiguity and distortion in terms of which various sorts of impurities arise, and adapt and develop frameworks through which they might be systematically assessed and disclosed. The rationale is to meet a perceived long standing and still growing need to incorporate appropriate qualifiers into quantitative information, this in turn arising from the intrinsically problematic aim of producing robust quantitative measures and estimates from many fields of geographical inquiry, and from the inherently impure conditions of their use. The frameworks discussed provide for protection against the conferment of bogus authority on inherently uncertain and pathologically distorted numerical measures and estimates, while at the same time facilitating due utility to be derived from whatever measures and estimates can be produced. Illustrations of the applicability and utility of the frameworks will be drawn from three different fields of geographical inquiry - environmental risk, energy policy and retail expenditure flow forecasting. Comment will be given about the wider applicability and limitations of the ideas formalised in this paper, and about areas for further development. While the paper is written with explicit reference to geography, its relevance to related fields, including environmental science, planning, economics, social policy, and many others, should be immediately evident.

KEY-WORDS: Quantitative geography, uncertainty, ambiguity, distortion, quality assessment, systematic frameworks.

Paper prepared for the 5th International Colloquium on Quantitative Theoretical Geography, Bardonecchia, Italy, September 1987.

1. INTRODUCTION

The process of quantification - whether a matter of instrumented measurement, mathematical and computer modelling, statistical inference, counting, probability calculus, extrapoloation, interpolation, intuition, guesswork, or whatever - can involve substantial complexities and impurities. The use of resulting estimates can be pathologically distorted. It is the premise of this paper that more can and should be done in terms of the recognition and control of these afflications. The aims of the paper are consequently: to take stock of key sources of uncertainty, ambiguity and distortion in quantification processes, and in the use of resulting estimates; and to adapt, develop and exemplify frameworks through which they might be more routinely and systematically assessed and disclosed.

2. RATIONALE

The following factors provide the paper's context and motivation:

* The fundamental role of quantificaton in the development of knowledge yet the persistent neglect within some fields of geographical inquiry, and in the use of the products of quantitative geographical research, of the many problems that the process entails.

* The seemingly insatiable demand for information in quantified form, not only in contexts in which unquestionably hard numerical estimates can be produced, but also in contexts where a lack of empirical robustness makes the production of quantitatively expressed information inherently problematic. Tendencies contriving to accelerate an excessive

demand for quantification are manifold: the increasing use of artificial intelligence, expert systems, and information systems - automatic pilots for the mediation of (what quality of?) quantitative information in ever greater amount; the possibility of an enthusiasm for quantitative techniques to outpace the inherent measurability of what is being worked on; and misguided beliefs in the automatic authority of quantitative representations as compared either to alternative forms of expression (verbal, pictoral, symbolic languages), or to sober hesitation in the face of what is not known.

* The need to facilitate the proper interpretation of quantitative knowledge. On the one hand, there is a need to guard against the conferment of meaningless precision, or bogus authority, on the results of inherently uncertain quantification processes. On the other hand, there is a need to guard against unfounded criticisms (and, in extreme cases, blanket rejection) of quantitative knowledge, regardless of its legitimate claim to greater clarity, rigour, precision, or 'truth' over verbal or other forms of expression. To put it another way, although quantitative expressions themselves may be of a superficially homogeneous form - one or a set of numbers - the differing sources and degrees of uncertainty and ambiguity in their underlying derivation renders their actual 'quality' intrinsically variable: high in some cases, low in others. As a matter of practical importance, anyone being provided with quantitative geographical knowledge as a basis for their own actions - planners, decisions makers, academics, members of the public, or anyone else - should be given due indication of the nature of known impurities, and therefore of the quality of resulting products, so that they can interpret them properly.

* The need to protect geography and its sub-disciplines against criticism (or even ridicule) when the products of earlier quantification processes are rendered obsolete. If at the time they had been guarded against overconfidence in their interpretation then accusations of overselling would have been preempted. In a period when geography, as with other disciplines, is coming under attack for extraneous reasons as well as for valid criticisms, such providence in the transfer of information may be very important indeed.

* The scope for heightened awareness of the various impurities and distortions which exist in the contexts in which quantitative information is used. These contextual impurities and distortions arise in part as a function of political or strategic uses of quantitative information, in part as a function of constraints on access to such information, and in part as a function of limited technical and interpretative competence of various interested individuals and organisations.

* The intellectual benefits resulting from an appreciation of broad structural similarities in the types of uncertainties, ambiguities and impurities in quantification that arise in ostensibly different geographical contexts: the same 'levels' (generic types) of uncertainty, ambiguity and impurity appear in different guises in different contexts. A sharper appreciation of these different 'levels' (or generic types) will make for improved intellectual capability, with benefits both of pedagogical and diagnostical kinds, and also in terms of new insights in comparative work - whether across different fields, or against the earlier state of knowledge within the heritage of a given field.

* The newly presented opportunity for making real headway in the identification and communication of generic classes of uncertainty and ambiguity in quantification. This opportunity arises partly from the timely recent appearance of new conceptual tools which open up the possibility of a hitherto unsighted level of formal meta-analysis of such problems. Various existing formal provisions (the assignment of confidence limits, error bars, or upper and lower bounds; the use of scaling techniques, and various forms of sensitivity analyses) can thereby be put into due perspective, and aspects of uncertainty, ambiguity and impurity not accommodated by such existing provisions can be acknowledged. Herein lie some real implications for the limits of existing formal approaches to 'uncertainty' analysis (eg for the confidence with which confidence limits can be assigned). Opportunity for wide appreciation of such a 'meta-level' of analysis of quantitative imperfections is securable now due to the quickening interest within the geographical community in the philosophical foundations of its own endeavours, and consequent developments and refinements in the appreciation of the nature of knowledge - whether expressed in quantified, verbal or other form - of the limits to what is known, and of deficiencies.

* The intellectual challenge of exploring the limits of formally disciplining the expression of uncertainties, ambiguities and impurities afflicting quantitative knowledege, and thereby of the relative degree of robustness of that knowledge. In effect, it is to address the challenge of developing and refining an appropriate formal language for representing related aspects.

3. SOURCES OF IMPERFECTION

With the possible exception of straightforward counting, all quantification processes are subject to uncertainty (error, inaccuracy or chance, for example), and many also to ambiguity (less than absolute definition). The reason for this lies in their inherently indirect nature, whether of physical or of social phenomena. There are key differences between different contexts, but these are not so much a matter of whether or not uncertainties and ambiguities exist, but rather of their source, degree and significance.

In terms of the production (as distinct from the use) of quantitative knowledge, some sources of uncertainty and ambiguity may be reducible by refinement and revision with the inevitable advance in the frontiers of research over time, and in the development of individual skills, techniques and methodologies. Some may be deemed negligible in the context for which the quantification is actually to be used. But there remain many further sources of imprecision and inexactness in quantification processes, and in many cases less than complete understanding of underlying phenomena which prevent the elimination of all significant sources of uncertainty and ambiguity. In this geography is hardly exceptional among academic disciplines. What is needed is positive assertion of fallibility and ignorance on related issues. Valid quantification cannot arise by attempting to hide such ignorance and fallibility.

In terms of the use (as distinct from the production) of quantitative knowledge, people's different interpretative competence, problem definitions, and strategic interests make for pathological distortion and

questionable legitimacy.

We will take stock below of sources of imperfection and impurity in the production and use of quantitative geographical knowledge under two broad headings: (i) measurement uncertainties (this relating primarily to how 'true' quantification is to the reality supposedly represented); (ii) contextual distortions (this relating primarily to impurities in the use of quantitative information). The two aspects are not completely separate, but their distinction is sufficiently emphatic to be deemed useful for present purposes.

3.1 THE POTENTIAL FOR MEASUREMENT UNCERTAINTY

Five sources of measurement uncertainty will be distinguished: validity of representation; technical error; human error; queries beyond.

VALIDITY OF REPRESENTATION: The question of the validity of a given quantitative representation in relation to what it purports to represent arises because there is often no direct (fundamental) measure for phenomena of interest, so that an indirect measure has to be used. Well known examples include the use of indicator species in ecology; the use of indirect indicators of economic and social phenomena; the spiking of laboratory samples to infer 'untraceable' elements; the use of sampling to infer characteristics of a larger (unobservable) population; the use of available (rather than desirable) levels of aggregation or resolution in statistics on employment trends, measures of pollutant levels, spatial data bases (*1), and so on; the use of results of laboratory experiments

as surrogates for 'real world' observations; the use of composite surrogates through techniques such as multi-dimensional scaling. In such cases it is desirable (though not always possible) to know how well the chosen indicator represents what it is being used to depict (*2). Indicators of crime, health care need, university excellence, de-industrialisation, agricultural productivity, quality of life, toxic and radiological hazard, and 'misinformation' in certain commercial data banks are recurrently problematic in this respect. If there are competing (indirect) measures to choose from, then a variable mix of pragmatic, political and epistemological considerations will need to be brought to bear, with an outcome that inevitably falls short of ultimate neutrality or objectivity. There are also the well known (though persistently uncontrollable) problems of observer bias distorting observation of the social (life) world, and therefore calling into question the validity of resulting representations; for example, of people's positions in attitudinal studies. We simply cannot necessarily define a single objective reality from the multiplicity of intersubjective perspectives that constitute society.

Some of the most persistent neglect of representational problems arises in the course of blanket application of formal techniques (*3) whose surface appearance of accommodation of social attributes (people's wants, preferences, problem definitions) masks their deeper neglect of the true subtlety and ambiguity of these very phenomena.

SCALE: From the point of view of identifying sources of imperfection in quantification processes, the significance of the scale (unit) of measurement is twofold. First, choice of scale (of 'object

discrimination') determines the degree of precision afforded: the finer the scale of measurement, the greater the degree of precision possible (grammes cf kilo; parts per billion cf parts per million; seven significant digits cf two significan digits; finer cf coarser zoning systems; more cf less disaggregated specifications of economic sectors or of populaton characteristics; and so on). To adopt a particular scale (or unit) implies 'rounding' up or down of anything at a finer level of resolution, for to allow finer detail would, in effect, be to adopt a finer unit. To this extent there is necessarily a degree of imprecision in all quantification. Good practice is to adopt a scale or unit of measurement for which the degree of imprecision will be tolerable, or negligible. Problems arise in the many cases where this is simply not practical (this complements the sort of points raised in consideration of validity of representation: see also again (*1)).

Second, to adopt a scale or unit of measurement which is inherently too demanding (fine) for the nature of the measurement or quantification actually being undertaken, or to manipulate measures on nominal or ordinal scales as if they were derived on interval or ratio scales, will merely lead to a misleading or illegitimate impression of precision. This is obviously a practice to be avoided by anyone genuinely interested in valid measurement. Rogue examples include the derivation or publication of statistics precise to five or six significant digits when many of the source statistics were more coarsely specified; and soil pH measurements given to three or four decimal places derived from an instrument calibrated to a single place. A related problem arises in attempts to make comparisons on the basis of 'different' units of measurement.

TECHNICAL ERROR: As the name implies, technical errors reside within the formal techniques (observation instruments) through which quantitative measures are derived. They include: constant and systematic errors of technical measurement instruments (lense distortion in aerial cameras, atmospheric dust distorting optical and electromagnetic measurements, temperature change altering the length of a physical measures); random and systematic (eg spatial autocorrelation) errors in statistical analysis; deficiencies in specification or calibration of mathematical models (in terms of overall fit, and in terms of specific refinements).

There is an important relationship between 'technical error' and 'scale', for technical errors which are within the margins of distortion already allowed for in the scale of measurement suited to a given context can be safely neglected. For those that are not, it should be a mtter of 'normal practice' to incorporate appropriate correction factors, or specify error bars, confidence margins, or whatever, in order to make due acknowledgement of them. These are automatically given in the case of some (though not all) techniques and statistical packages, though there are different degrees of thoroughness with which this can be done, and not necessarily a unique choice of methodology.

HUMAN ERROR: Human fallibility may generate error in a number of ways. A record may be incorrectly copied, sloppy fieldwork may led to defective samplings and readings, misuse of equipment may result in malfunction, typing errors may result in faulty data processing, fragmented hierarchy may lead to orders not being followed, 'historical' data may contain 'forgotten' flaws, lack of technical competence may lead to rogue quantification, or rogue model specification and use, absence of common

meaning may lead to the 'wrong' thing being measured, correction factors may be misapplied, time constraints may lead to rushed work, observer bias may (will) distort participant response. Mistakes and blunders, as well as more subtle types of error, are commonplace, and not always easy to detect. The qualifications and track records of those undertaking the work may be reasonable indicators in some (by no means all) cases; and repeated, independent doubling may identify sources or error in other cases (though how many actual publications contain sufficient detail for independent verification by the reviewer or reader?).

QUERIES BEYOND: A technique may be applied correctly to produce a quantitative estimate for some phenomenon of interest, but there may still be legitimate ground for doubting the veracity of the product. A statistical inference may be given at the 99% level of confidence, but we may be much less than 99% confident that it is 'right', because of the possibility of sources of uncertainty beyond those that the methodology itself can encompass. Our theoretical understanding of what is being estimated may not be established, strong and refined, but speculative, weak and immature. This may lead to uncertainties over model specification (such as whether linear of log linear forms should be used for regression, or as whether cause-effect relationships established under laboratory conditions can be assumed to hold in 'real world' contexts), and model calibration. There may also be crucial areas of ignorance about stimuli whose origin is external to a chosen system of interest (unforseeable price shifts in energy markets, for example) or as yet 'unknown' (the ozone depletion problem, for example). Such queries are characteristic of embryonic, immature or ineffective fields of inquiry (as distinct from well developed, mature fields).

3.2 THE POTENTIAL FOR CONTEXTUAL DISTORTION

We attempt to take stock of contextual distortions - impurities which vitiate universally 'free' access to and 'pure' use of quantitative information - under three headings here: appropriateness; strategic bias; and performance. At this stage, these headings are neither definitive nor mutually exclusive, but merely starting points for a partial organisation of wide ranging and open issues.

APPROPRIATENESS: This heading is used as a basis for questioning the relevance of quantitative information to a given problem or issue of interest. Just as lack of fundamental measure or absolute definition can confound the production of valid estimates for phenomena of interest (as referred to above under 'validity of representation'), so also can these problems confound their use, by rendering ambiguous the relevance of quantitative information to a given problem. The relevance, or appropriateness, of quantitative information to a given context can also be undermined on account of what can be referred to as 'openness of meaning'. Many a controversy and sterile debate has arisen in geography (and elsewhere, not least overtly political contexts) because of different meanings an interpretations for what is being referred to. Terminology can have a considerable degree of fuzziness, so that clear clear definitions cannot be secured; other terms have great variation in what they are taken to mean (rational man, environment, region, crime, unemployment, de-industrialisation); the natural meaning of other terms may change as new developments take place. Such problems of openness of meaning seriously confound the interpretation of agreed statistics on related

issues - crime rate, unemployment rate, quality of life, poverty, and, to a greater or lesser extent, indicators for virtually any other social phenomena one might care to name. If we seek to override the inherent ambiguity in meaning by imposing a precise definition, the temporary and context specific nature of that definition must be fully recognised. With time, or new context, or participation of new social groups in related debate, it may need to be re-built, or at least re-examined in order to assure ourselves that it is satisfactory. Recent attempts to override inherent ambiguity of definition (such as strict set membership rules, Atkin 1974, (*4), though successful in some contexts, cannot claim general applicability, still less widespread acceptance (Couclelis 1983, Macgill 1983). Morover, as Kaplan warns of attempts to overcome inherent ambiguity of meaning:

"The demand for exactness of meaning and for precise definition of terms can easily have a pernicious effect as I believe it often has had in behavioural science. It results in what has been aptly named the premature closure of our ideas" (Kaplan 1964 pp 62-71, cited by Harvey 1969, p304).

STRATEGIC UNEVENNESS: A further class of considerations has to do with differing degrees of social good derivable in the use of quantitative information. Ideally, such information should heighten awareness and knowledge among all concerned. More realistically, it is likely to be used to favour the interests of some agencies, to the detriment of those with inferior access and power (due to restrictions in information flow, lack of technical know how, or other factors). In other words, there is unlikely to be an ideal exchange or transfer of quantitative information, but rather what we might call instrumental or strategic unevenness in its use. People's interpretative competence - their ability to understand the origin and process of producing quantitative estimates - can also severely limit their use of related information.

CONSEQUENCE: A somewhat different aspect of use has to do with the consequence or significance of a given quantitative estimate for a particular phenomena in relation to notions of 'performance', 'accomplishment', 'moment' or 'criticality'. In the public policy field, for example, it is necessary to have good contextual knowledge in order to be able to assess whether the numerical value of an indicator is de facto evidence of 'good' or of 'poor' performance. One can easily be misled by the sheer magnitude of numbers (such as when people believe radiation measurements given in becquerels represent greater 'hazard' than measurements given in curies). Correspondingly, in fields of environmental measurement, for example, indicators can often only be sensibly interpreted if their relation to critical, target or threshold values is known.

4. CRITERIA FOR QUALITY ASSESSMENT

We draw on the foregoing discussion of ambiguity, uncertainty and use values and assemble here elements of formal frameworks to facilitate more routine and systematic acknowledgement of uncertainties and ambiguities in quantification processes, and of impurities in the use of resulting products. Extending the pioneering work of Funtowicz and Ravetz (1987), frameworks are to be composed consisting of a number of individual criteria through which the validity of quantitative estimates might be assessed, together with a range of descriptive modes for each criterion in terms of which a given quantitative estimate may be judged (*5). Suggested criteria for these frameworks, the outcome of a mapping of some of the

different sources of ambiguity, uncertainty and impurity identified above on to convenient analytical categories (*5a), are as follows:

PRECISION	RELEVANCE
ACCURACY	CONSEQUENCE
ROBUSTNESS	PURCHASE
DEFINITIONAL VALIDITY	EASE OF USE
DATA QUALITY	COMPUTATIONAL DEMANDS
PROCEDURAL COHESION	FLEXIBILITY
THEORETICAL DEVELOPMENT	
STATE OF THE ART	
EXTENT OF REVIEW	
OUTCOME OF REVIEW	

This cannot claim to be an exhaustive set, but rather an open ended list of significant counts in terms of which the author has been aware of significant variability, often deficiency, in quality in a number of contexts.

The range of descriptive modes to be given for each criterion are a way of characterising different possible levels of distinction, or of weakness of a given quantitative estimate in relation to individual criteria. For example, for the criteria of 'precision' and 'robustness' we can have, in decreasing order of distinction:

PRECISION: superlative, good, moot, remis, reckless

ROBUSTNESS: strong, resilient, elastic, weak, wild

The idea in presenting modes for criteria in such a formal ranking as this is to encourage an appropriately explicit assessment of quality in terms of chosen criteria.

The full set of criteria considered in the present paper, and suggested descriptive modes, are given table 1. The criteria, listed at the left hand side of the table, encompass technical (rows 1 - 3), methodological (rows 4 - 7) and epistemological (rows 8 - 10) aspects of the production of quantitative knowledge (ie measurement uncertainties), and functional (rows 11 - 13) and practical (rows 14 - 16) aspects of its use (ie contextual distortions). Not all of the criteria will be 'required' in order to arrive at an adequate assessment of quality in all contexts (though their range and number - even in this 'incomplete list' - is indicative of the myriad pitfalls and distortions of quantification processes). A framework comprising three or four may well be sufficient. Selection will be context dependent - on the nature of the quantitative process and what the product is to be used for. The quantitative information which is to be assessed in terms of resulting frameworks may be given as a single number - fraction, decimal, integer or whatever, a set or matrix of numbers, a range (with or without confidence intervals, and based on greater or lesser extents of sensitivity analysis), and may be model based, measured, counted, or whatever.

The first three criteria enable an assessment of the technical quality of

the quantity (or set of quantities) under consideration. The first provides a scale for describing the deemed precision of a quantitative estimate. Given the different degrees of specificity in which quantitative estimates can be given (point estimates, two, eight or whatever significant digits, probability estimates, intervals, upper limits, and so on, as well as different levels of resolution in discriminating subjects of interest) then it is only sensible to ask whether the precision given is what it claims to be, or whether point estimates have been given when ranges or intervals would have been better, or whether seven significant digits have been given when there is justification for only two.

The second and third criteria are concerned with further issues of technical merit, reflecting possible aspects of technical (random or systematic) or human error which may undermine the accuracy (or reliability) of quantitative esimates (*6). In some cases, conventional goodness-of-fit statistics are (or can be) built into quantification processes, and the interpretation of resulting quantitative information can then be related to particular modes from row 2. In other cases, however, the question of accuracy cannot be answered conclusively, or even directly, either because of inability to 'observe' the reality directly (for example, in forecasting contexts), or because of lack of agreement about suitable terms in which comparisons with reality should be made - such difficulties are better acknowledged than ingored (notions of the natural predictive range of a model may afford it greater validity in some such contexts than others) and the different modes of 'accuracy' given in the second row can be correspondingly deployed. It is also worth noting a crucial trade-off here: a quantitative estimate given originally as a range may warrant a higher 'accuracy' rating than one given as a point

estimate, or a narrower range, for the former has more scope for spanning the 'true' value.

There are the related matters of robustness and confidence (row 3) - of whether the quantitative estimate is resilient to changes in data inputs, parameter values, mathematical model specification or other sources of variation in the quantification process? Sensitivity analysis can test this to some extent, often ranging to probe the existence and impact of critical values, and framing answers in formal probability terms (*7). Where sensitivity analysis has not been undertaken, one may wish to judge a quantitative estimate rather differently from where is has.

The next four criteria enable an assessment of methodological aspects of the production of the quantity (or quantities) under consideration. The fourth criterion is a means through which to query the 'units of measurement' observed in the process of producing quantitative indicators (*8).

Considerations of validity of representation above cast doubt on the 'truth' of quantitative estimates which are 'indirect' or over aggregated measures of phenomena of interest. Related problems can be further compounded given scope for 'openness of meaning': compilers of data sources may well simply 'count' or 'measure' what is obvious to their own common sense (but with no guarantee of uniformity between different compilers) so, without further qualification, no guarantee of consistency can be assumed. Five possible modes are given under the heading of 'definitional validity' in order to encompass such considerations. Fundamental (or primary) refers to those cases where the definitions used are fully attuned to the context in hand - it arises much less than might

casually be thought. Fundamental physical measures of length, mass, time, when used directly, are cases. Standard refers to those cases where the definitions used are those of some established practice - for example standardised zoning systems - a coarser approach than using individually tailored units. The mode 'convenience' is for cases when the definitions used are not necessarily the most appropriate but have been assumed to be convenient ones for the purpose (for example, on the grounds of ready availability, cost or manpower, or use of ad-hoc calibration methods and 'spiking', when the phenomena of interest cannot be observed directly). While 'symbolism' could also be construed as a form of convenience, it is used here to denote cases where, for example, considerations of institutional legitimacy or prestige dominate the choice of units. The final mode 'inertia' characterises the situation where particular definitions continue to be used for traditional or historical reasons, regardless of their validity or utility.

Quantification processes involving a significant data gathering element - whether as the chief component of the process, or as an input to a further (for example, mathematical modelling) stage - call for different evaluation criteria, for example, in terms of the modes of the fifth criterion. (Note that souces of imperfection here operate at a different level than that implied in the way accuracy-robustness has been considered above and reflected in rows 2 and 3: factors of uncertainty beyond the reach of, but potentially undermining the veracity of, even the most exhaustive sensitivity analysis.) The quality of data collction can be extremely variable, ranging from reliable primary data of controlled laboratory standard, or as compiled by a first rate task force, to secondary data of lesser quality - including proxy measures and sheer

guesswork. Trade-offs between possible modes of this criterion and possible modes of some of the other criteria are also worth noting: for example, the demands for policy relevance, or the data demands imposed by a particular theoretical model structure, can in some cases only be met if there is a 'sacrifice' of data quality ideals.

The criterion of 'procedural cohesion' is included for contexts in which it is desirable to acknowlege the degree of unity between different elements of a multi-stage quantification process (*9). For example, a model based forecast of future population size will involve at least three stages: formulating the model; securing the data to be fed into it; and running the model. Correspondingly, a census of household characteristics across different national or regional boundaries will entail at least the two stages of enumerating the required characteristics for individual tracts, and synthesising the outcomes from the different tracts. Wherever multiplie stages arise, there is a need for awareness of how well they interlock. At best there will be full dovetailing of different stages (for example, full negotiation between different agencies involved); at worst, fragmentation and divergence.

The criterion of 'theoretical development' would be of interest where quantitative estimates are derived from conceptual or mathematical models. Depending on the degree of understanding of the 'real world' mechanisms the model is designed to depict, its quality might be suitably assessed in terms of whether its driving mechanisms are based on laws, theories, hypotheses, or whatever - see row eight (*10). Of course, since such theoretical considerations can affect accuracy, and have implications for data requirements, the lack of complete independence between criteria

is again apparent. For example, with care it is possible to get good precision from 'spiking', but in the absence of traceability, accuracy can be less than certain and systematic errors can arise. Correspondingly, theoretically sophisticated model specifications may make unrealistic data demands.

Criteria 8, 9 and 10 refer to epistemological aspects of the development of knowledge. These foundations to potential truth claims operate at a 'deeper' level than the criteria considered so far, and provide a perspective against which to assess earlier aspects. The eighth row sets out (in Kuhnian-type terms) what can be expected in the light of the state-of-the-art of a given field. One cannot expect to find well tested theories in an embryonic field, and may need some convincing argument to tolerate mere speculation from an advanced field. In general, then, some broad indication of the state of development of the field itself may well be in order, so as to put entries from row 7 into relief, or in some cases to take their place, and the modes given in row 8 span an appropriate range of possibilities (*11).

Epistemological considerations also embrace more overtly sociological dimensions, notably the extent of peer review of the quantification process and the outcome of that review (rows 9 and 10), following the philosophy that the validity (truth claim) of any 'knowledge' can only ultimately be assessed via discourse (and ultimately through consensus). Since legitimate assessment of virtually any of the preceding counts demands a degree of review within some peer community (*12), then the extent of this, and the outcome, can be usefully acknowledged. The review undertaken to verify or validate a quantitative estimate may be limited to

self appraisal (low), or may extend quite widely to independent verification within a full peer community (*13).

Criteria 11, 12 and 13 move from the production of quantitative estimates to overtly functional aspects of their use. The eleventh criterion invites an assessment of the actual relevance of a quantitative estimate to the 'real world' problem to which it ostensibly relates (c.f. appropriateness of 'definitions' under criterion 4.) Such considerations are closely related to those discussed under 'openness of meaning' above. As is widely appreciated, model resolutions can be frustratingly deficient - models valid only for short term projections being called on to produce long term scenarios; highly aggregated generalised models being used for specific inferences; serious mismatches between questions policy makers want to address and issues which models can articulate; spatial zoning systems which are much coarser than desirable (national for regional; regional for local *14); there can also be ambiguity and lack of consensus over what the appropriate measures or indicators for a given problem actually are; invalid transfer of models from one context to another (for example, laboratory results to 'real world' contexts). There is also the question of choosing between possible alternative means for producing quantitative information for a given purpose. The different entries in the eleventh row do not encompass all of these considerations, but perhaps makes some useful headway in terms of their possible representation, under the label of 'relevance'.

The criterion of 'consequence' is a basis for assessing the relative significance, or degree of achievement, reflected in a numerical estimate or quantitative indicator. For example, there are many instances in public

policy fields where numerical indicators are used as explicit indicators of 'performance'. A suitable scale for such an assessment might run from 'distinction' (greater achievement could hardly be expected) through successively less commendable descriptions, to a position warranting some derision about the prevailing state of affairs. The five modes in line 12 are suggested accordingly in descending order of merit. Modified interpretations of such descriptors, (or modified descriptors) could be used for assessment of measurement criticality in relation to some given standard. More so than for other criteria given so far, numerical indicators (or other quantitative estimates) will need to be located within their broader contexts when being considered in relation to such modes.

'Purchase' acknowledges the possible unevenness in strategic value of quantitative information. This is a crucial issues, though problematic to represent, and perhaps the least satisfactory in terms of the criteria given in this paper. What is given in row 13 is a somewhat crude condensation of the terms in which the functional use of knowledge has been discussed elsewhere in the literature (Habermas 1972, Choi 1987). The 'ideal' position is for quantitative information to serve emancipatory interests - the shape of quantitative knowledge base mirroring public interests and concerns and serving a foundation for true consensus. More realistically (and less desirably from the public interest point of view (*15)) quantitative information will be used to advance the interest of specific private groups. It can be used as 'currency' in political debate; or restricted access and its presentation in specialist elite languages can debar the development of broad social understanding. The difficulties of making appropriate assessments, and the time dependency of any

assessments made are among the formidable difficulties associated with this criterion. However, it is perhaps better to include the category in some form, than to exclude it completely, and a range of possible modes are accordingly suggested for the criterion of 'purchase'.

The final criteria have to do with practical aspects: ease of use (how much is demanded of people's minds); computational demands (how much is demanded of their hardware); and flexibility and adaptability. Again, interrelationships between criteria can be important to consider: a userfriendly computational package (scoring well on row 14) which has been subject to only limited review (scoring poorly on row 9) should be viewed with some scepticism.

5. FRAMEWORKS FOR QUALITY ASSESSMENT AND DISCLOSURE

The sixteen criteria described above by no means exhaust the terms in which the quality of quantification processes, and the use of resulting products, might be assessed. However, they will be accepted here as sufficiently extensive to provide an adequate basis for present purposes. As already remarked, not all are relevant to all contexts. What is needed is consideration of the history of production of a given estimate and its context of use, and corresponding identification of those criteria which are most relevant to these considerations.

Frameworks for assessing and disclosing the quality of any given quantification process can be constructed from criteria of the above kind by selecting those that are specifically relevant to the chosen context, and 'scoring' in terms of the normative range of modes. For example, evaluation of the products of a given modelling initiative in the urban

and regional field might suitably proceed on the basis of criteria 2,5,7 8,9,10, and 11. By virtue of the normative ordering among the entries which constitute the scales for each criterion (successively 'better' as one moves from right to left along a row), scores can be used as a concise way of representing individual elements of each row. For the example of an urban and regional modelling initiative, the string of scores - one for each of the chosen criteria, in order - (2,2,1,2,4,1,4) would denote a modelling initiative whose reliability is open to doubt (a score of 2 from the second), whose data is 'calculated' (a score of 2 from row 5), whose theory is best described as 'hypothetical' (a score of 1 from row 7), whose parent field is as yet 'intermediate' in its state of development (a score of 2 from row 8), which has had wide review (a score of 4 from row 9), but low acceptance (a score of 1 from row 10), and which is directly relevant to a given problem (a score of 4 from row 11) . Rather a mixed effort, and well short of the 'ideal' string (maximum distinction on all of the chosen counts) of (4,4,4,4,4,4,4).

It may well be impossible to attain the ideal string of top scores in many, indeed most, contexts. The point is not to agitate for fulfilment of utopian expectations. Rather, in realistic acknowledgement of the many possible impurities and distortions in quantification, quality assessment frameworks of the kind assembled above can be used to provide guidelines for critical though and subsequently invoked as vehicles for disclosing what is being achieved in relation to some 'ideal type'. Their use should facilitate wider appreciation of the real significance of a given quantitative estimate, and fix more firmly an understanding of what is being offered. In particular: such frameworks:

* Deriving the scores for any given quantitative estimate in

relation to chosen criteria should be undertaken as a process of critical reflection; assessed quality can subsequently be communicated to others.

* Comparisons between assessments for different quantitative estimates could be very insightful, setting the quality characteristics of individual estimates into relief against others, and acknowledging also that different quantitative estimates, explicitly or implicitly, claim distinction on different counts.

* Claims for quantification must be in line with (the legitimately assessable quality of) what can be delivered. Rhetoric that is out of line with formal evaluation should be rejected. Over-enthusiastic claims for quantification, and unduly damning criticisms each in their own way inadvertently circumvent this 'test'.

* Poorer performance than is 'necessary' on any of the given counts should be rejected, whether manifest in terms of poor data, impoverished theory, or whatever.

* There should, however, be no shame about a less distinguished performance than is possible on any of the given counts. One should not expect delivery of the impossible. Low scores on certain counts may be a legitimate reflection of reality, not of inherently poor quality of endeavour on the part of those involved.

* Weaknesses on certain aspects may confound the attainment of distinguished performance on others: for example, irremediable data deficiencies, or endeavour in a field which is as yet embryonic in its development.

* 'Unusual' strings of scores should be viewed as possible pointers

of something suspicious - claims for strong theory in embryonic field, for example.

* By choosing the most 'testing' (least redundant) criteria for any assessment, then (and only then) the framework can be used as a reliable 'cautionary' signalling system. Such a conservative approach should avert any false claim to distinction in the production and use of quantitative information.

6. THREE ILLUSTRATIVE EXAMPLES

Three examples illustrating the applicability of quality assessment frameworks based on some of the criteria given in table 1 are summarised below. The examples span a range of different fields of geographical inquiry and contrast in terms of the criteria brought to bear. The first example concerns an estimate derived from modelling the transmission and transformation of radionuclides through the physical environment and to target cells in the human body. It entails issues of technical complexity and scientific uncertainty, and originally arose in a context of significant political sensitivity. The second, in contrast, deals with the production and use of performance indicators in a topical public policy context - official energy efficiency campaigning in the UK. Issues of interest here include the institutionalised compilation of particular statistics, and the purchase sought through their publicity. The third example, and in further contrast to the first two, concerns the development and successive refinement of generations of models in the urban and regional field for forecasting retail expenditure flows between consumers and shopping centres. It is again, in a sense, 'representative' of a wider class of modelling initiatives in other contexts.

6.1 EXAMPLE 1: THE ESTIMATION OF RADIATION EXPOSURES FROM A SPENT NUCLEAR FUEL REPROCESSING INSTALLATION

There have been growing demands in recent years for the production of 'hard' quantitative information about the possible malign health effects of civilian nuclear power development. However, uncertainties and gaps in knowledge in the relevant scientific fields - radiobiology and epidemiology - render problematic the production of such information. A specific case study illustrates well this state of affairs.

In the report of a recent official inquiry, it was estimated that at most 0.1 of a death from leukaemia to the under 20 year old local population could have been caused by discharges from the nuclear installation at Sellafield, West Cumbria, UK over the period 1945-75 (Black 1984). The smallness of this quantitative estimate of 0.1 was a crucial basis for a subsequent official reassurance to people concerned about a possible health hazard in the vicinity of that installation. We show here how a quality assessment framework based on four of the criteria summarised above - 5, data collection; 7, theoretical development; 8, state of the art; and 10, outcome of review - can be deployed to (re)couple the numerical estimate of 0.1 to crucial qualifying considerations. These particular criteria are chosen for the quality assessment here in order to bring out characteristics of measurement uncertainty beyond those which were substantially addressed in the original estimation procedure (*16).

The estimate of 0.1 was derived from radiobiological studies involving three key stages: (1) reported discharges from the installation, routine monitoring data, and models of the transmission and transformation of radiation through the environment; (2) the uptake of radionuclide contamination from the environment giving doses received by young people in the local area; and (3) models of the pathogenic effects of these

estimated doses. Quality assessment codings for each of these three stages are given below, these codings in turn deriving from specifically designed interviews with relevant specialists (for more details, see Macgill and Funtowicz 1987).

The data input to the first stage was regarded to be of variable quality - partly historical field data, partly interpolation between measurements, and partly educated guesswork (scores of 1 - 3 for data collection). The processes whereby radionuclides in the discharges may pass through various pathways in the environment were described as being understood in broad theoretical terms; theoretically based models are used to depict these processes. A score of 3 on criterion 7 would reflect this. The degree of peer acceptance of the estimate was regarded to be 'medium to high' (scoring 2-3 on criterion 10), with the integrity of the stat of the art disputed only by rebels (score 3).

Overall, the string of quality assessment scores (see line 1, table 2) depicts a weighted qualification about the possible integrity of the estimation process; its data input aspect being particularly weak. Overall it conveys a picture neither of total confidence in understanding, nor of complete ignorance about the underlying phenomena of interest in stage 1.

For the second stage, the estimation of the uptake of possible radiation contamination from the environment to various sites in the human body, the data input was again described as being a mix of field data (monitoring), interpolation, and educated (or, someone suggested uneducated) guesswork (scores of 1-3). The estimation itself was again said to be derived from theoretically based models (score 3). The epistemological development of the field itself was regarded as being more questionable than for stage (1), with competing theories on a number of crucial aspects (scoring 2).

The peer acceptance of the result was regarded as being 'medium to high' (score 2-3). Overall the assessment (see line 2 table 2) depicts a somewhat more heavily qualified result than for stage 1, given the loss in experts' (technical) consensus about the state of maturity of the field.

Stage (3), estimating the biological effects of radiation doses was generally regarded as being less secure a process than that of the previous two stages. Knowledge in this field has been significantly developed in laboratory and military contexts, but there were regarded to be great problems in producing reliable estimates for man, for low level discharges, and still greater problems for a specific cancer for children. The data aspect was said to comprise some field data, and some educated guesswork, or worse (score 1-3). For the theoretical aspect, the modes statistical processing and computational models were deemend the most fitting categories (scores 1-2). The peer acceptance was described as 'meduim to low' (a score of 1-2); and there was a view by some interviewees that the field is characterised by 'competing schools' and by others that it is as yet embryonic (score 1-2). Overall, the derived quality assessment scores here (line 3, table 2) depict a marked weakness in this stage of the estimation process.

All in all, the contrast in the derived quality assessment codings for each of the three different stages reflects the differing levels of understanding of the processes which constitute each stage and, in turn, brings out crucial differences in the degree of qualification which should be associated with any corresponding quantitative estimates. The contrast is more striking if we follow the convention of choosing the weakest of possible alternative codings for any given mode: this then gives, for stage (1) (3,1,2,3), for stage (2) (3,1,2,2), and for stage (3) (1,1,1,1).

The overall implication of the above sequence of quality assessment codings - far from the (unattainable) ideal of (4,4,4,4) is that the 'true' estimate of radiation induced health risk as a result of discharges from the nuclear installation in question may not be 0.1 deaths over the given period (contrary to initial impressions), for we could only have a high degree of confidence that this were the 'true' number if the resulting pedigree codings had been higher. At the same time, the low scorings, particularly for the last of the three stages, means that it is not possible to derive any other estimate than 0.1; the 'true' figure may be much lower, or indeed much higher.

It is worth stressing that the appearance of a relatively low quality assessment scorings should be occasion neither for shame nor for concealment in terms of quantification per se (though the political ramifications might prove problematic). It is merely a true reflection of the quality of what c has been produced in the face of irremediable difficulties, uncertainties, and gaps in scientific knowledge, and within the limits of timescale and resource inputs to the studies through which the estimate was produced. In this respect it is within the realm of what Weinberg (1972) has termed 'trans-science', where a problem can be formulated as a subject of scientific study, but is beyond the capability of scientific study to deliver a definitive result.

It would, of course, be possible to add further criteria from the earlier table to the four considered above, for example, explicity to acknowledge the deemed accuracy of the number 0.1 (criterion 2), or to reflect is political standing (criterion 12). As regards accuracy, we would point out that the four criteria actually used operate at a 'deeper' level than accuracy; and given the relatively undistinguished outcome on these four, accuracy is inevitably in doubt (so redundant as an additional criterion).

As regards 'purchase', then it is interesting to reflect that the 'smallness' of the number 0.1 perhaps affords it considerable political purchase, notwithstanding its questionability as 'hard' number (any number between 0.0001 and 10 may be just as plausible?).

6.2 EXAMPLE 2: PERFORMANCE INDICATORS OF OFFICIAL ENERGY EFFICIENCY CAMPAIGNS IN THE UK

Among 'facts' recently published by the UK Department of Energy about the need for greater energy efficiency and progress in the pursuit of related official campaigns (Department of Energy 1985) were the following:

- £7 billion a year is wasted by Britain through the inefficient use of energy
- over 1850 jobs have been created in local energy projects
- over 120,000 dwellings have been insulated by voluntary projects
- more than 5,500 energy managers have been appointed across the country

These 'facts' will be assessed here against five of the criteria listed in table 1, specifically: 2, accuracy; 4, definitions; 5, data collection; 11 relevance; and 12 accomplishment. Again, these are not the only criteria that might be appropriate, but they have been deemed those most relevant, and able to capture key aspects of the nature of the indicators and their wide publicity. The quality assessment scores given below are drawn from a fuller account elsewhere (Macgill and Sheldrick 1987).

The first indicator £7 billion wasted, achieves a score of 1-2 on accuracy (other organisations have produced somewhat different estimates for how much could be saved); a score or 3 on definitions (little room for errant discretion on the part of those producing the indicator); a score anywhere between 0 and 4 for data collection (because the indicator was

derived from several sources, including in depth studies, educated guesswork, and worse); a score of 4 on 'relevance' (because money is a direct measure of what the official energy efficiency campaign was focussed on); and a score of 1 or 2 on accomplishment (because the £7 billion figure does not represent a maximum economically and socially achievable target, still less a maximum theoretically achievable target). These scores are summarised in table 2.

The second indicator, 1850 jobs created, achieves somewhat different scores: 1-2 for accuracy (discrepancies in the survey used, and in its official interpretation); 2 for definitions (failure to distinguish whether the 'people employed' are full or part time, and failure to indicate the nature of their employment); 2-3 for data collection (part direct survey, part extrapolation and interpolation); 2-3 for relevance ('jobs created' being at best an indirect measure of energy savings); and 1-2 for accomplishment (much more could be achieved in terms of establishing local energy projects).

The third indicator is based on less than concrete 'definitions' (the terms 'insulation' and 'dwellings' being given different meanings in different cases - score 1-2); suffers from the same sort of reliability and data collection shortcomings as the 'jobs created' indicator (score 1-2 on reliability and 2-3 on data collection); is again a 'convenient' (not direct) measure of energy saving (score 2-3 on relevance); and again constitutes less than maximum accomplishment (score 2 -3).

Finally, the energy manager measure can be assessed. The reliability of the figure of 5,500 is undermined by the patchiness of the survey from which it was derived (score 1-2 on accuracy; 1-3 on data collection); the definition of 'energy manager' is open to different interpretations (score

2); as regards 'relevance', although in some cases the appointment of an energy manager would indicate that a firm takes its energy costs seriously, in others it may be purely cosmetic (symbolic) (score 1-3 on relevance); the accomplishment is accordingly difficult to asses - in some cases energy managers have been deemed outstanding successes, in others (symbolic) plainly not (score 1-4).

All in all, the strings of scores for the four indicators show that they are far from being 'facts-which-speak-for-themselves'. Taken at face value, this 'finding' may well do no more than confirm many people's intuitive doubts about so-called 'facts' that get produced in public policy debates in a wide range of contexts - by no means only energy efficiency campaigns. Beyond this, though, the framework is revealing in disclosing the particular aspects on which individual indicators are particularly weak, and those on which they are stronger: there is manifestly considerable variability amongst the different indicators. While some sources of imperfection may be irremediable, being inherent in particular sorts of data collections and production processes, or in the concept of energy efficiency itself in the present context, it may be possible to act on identified sources of weakness in other cases, and develop refinements to avert needlessly low 'scoring' on some of the criteria. This may be a particularly important positive contribution to emerge from the type of critical quality assessment being advocated in this paper.

6.3 EXAMPLE 3. RETAIL EXPENDITURE FLOW FORECASTING

Aspects of the considerable development and refinement in model based estimation of retail expenditure flows will be described and evaluated here in terms of six of the criteria from table 1: 1, precision; 2,

accuracy; 5, data collection; 7, theoretical development; 8 state of the art; and 10, outcome of review. The models themselves are reviewed in, for example, Clarke (1986), Guy (1987) and Wilson (1986).

Case 1. For many years the most common method of forecasting retail expenditure flow used singly constrained spatial interaction models of the type:

$$S(i,j) = E(i)A(i)W(j)f(c(i,j))$$

where $A(i) = [\sum_k W(k)f(c(i,k))]^{-1}$

S(i,j) is the interaction (retail expenditure flow) between a residential zone i and a shopping zone or centre j; E(i) is some measure of generation of shopping expenditure from i; W(j) is some measure of attraction of shopping expenditure to j; c(i,j) is the cost of travel from i to j; and f(c(i,j)) is such that interaction cost increases lead to declining flows.

With $f(c(i,j)) = d(i,j)^{-2}$ we have what by today's standards (which is how this and other cases will be judged here) would be considered a very rudimentary model: rather coarse specification (moot precision - score 2); rather coarse assumptions (leading to doubtful accuracy - score 1); data requirements could be met by direct survey (score 3 on data collection); theoretical development is barely more than working definitions - particularly in the way distance has been handled - score 0); the latter a particularly low score in light of the current state of the art (intermediate or perhaps even advanced - score 2 or 3); outcome of review - a score of 1, say.

Case 2. Refinements over the years have seen much improved model

specifications, for example (Wilson 1986):

$$S(i,j,m,g) = A(i,m,g)E(i,m,g)W(j,m,g)f(c(i,j))$$

where $$A(i,m,g) = [\sum_k W(k,m,g)f(c(i,k))]^{-1}$$

The additional indices m and g are attached to the key variable in order to accommodate (m) different types of people - according to car ownership, household structure, household income, for example - and (g) different types of retail goods. This more detailed specification, along with more refined assumptions embedded in the model mechanism should be reflected in the quality assessment scores which are assigned. The more refined assumptions include better representation of the effects of competition between retailers in the attraction of consumers as a subtle combination of centre attractiveness and the travel cost function which relate these to distance from each residential zone; recognition of travel cost itself being though of as a generalised cost made up of a number of components such as out-of-pocket money cost, travel time, waiting time, parking charges, and so on; and a recognised theoretical foundation, such as entropy maximisation (see also Roy (1987) for the generation of the overall model specification).

Assessing this more refined formulation in terms of the same criteria as before, we must acknowledge much improved precision (score 3), though accuracy is perhaps still moot, because of the greater data demands (score 2); data requirements are more demanding, and perhaps needing a significant element of interpolation, and not just direct survey (*16) (score 2); theoretical development is much improved (score 3 or 2); in line with the state of the art (score 3 or 2); and the outcome of peer review would be expected to reflect the higher ratings on most criteria

(score 3). The balance of scores overall reflects significant improvement in estimation procedure between cases 1 and 2, though not for all criteria. (Intermediate stages of refinement in modelling approach - between cases 1 and 2 described above - would, of course, be expected to display more modest improvements in scoring.)

Case 3. Guy (1987) considers a hierarchy of increasingly complex shopping models based on generalised linear modelling procedures. For example (case 3a), a simple unconstrained spatial interaction model linearised to the form:

$$\ln S(i,j) = \beta_0 + \beta_1 \ln E(i) + \beta_2 \ln W(j) - \beta_3 \ln c(i,j)$$

with a restricted domain of legitimate applicability would be expected to score in a smilar way to case 1 (see scoring for case 3a in table 2). A Poission formulation (case 3a), on the other hand, entailing also refinements in terms of extra explanatory variables, origin specific distance decay parameters and competing destination terms

$$p(S(i,j) = K) = \frac{1}{k!} \exp(-\lambda(i,j))\lambda(k,i,j)$$

(where (i,j) is estmated from a form such as that for lnS(i,j) above) would be expected to score more like case 2 (see scoring for case 3b in table 2). Moreover, the more precise calibration claimed of the linearised form (cf case 2) might be expected to yield better accuracy (though perhaps the theoretical development is less advanced?)

Case 4. Openshaw (1987) is iconoclastic in his approach to spatial

interaction modelling. His advocacy of data base exploration modelling entails maximum utilisation of increasingly powerful computer capability to explore any number and form of model specifications, with the aim of choosing that with best overall fit to whatever data is available. This promotion of data-bound model accuracy at the expense of theory can lead to the selection of clumsily eccentric (albeit well fitting) model specifications, such as:

$$S(i,j) = \text{arc sin}\,(\exp(O(i)^{\beta})\sin\,(D(j)^{\gamma})\ \text{cotan}\,(A(i)^{\alpha}) + \text{erf}\int_{-\infty}^{+\infty} c(i,j)^{-\pi\phi}$$

Quality assessment scores should reflect the rather different style of such a model (see table 2, case 4).

Comparing with case 2, we have a similar degree of precision (score 3), improved accuracy and more sympathetic data requirements (scores of 3 on each of these criteria); theoretical development is self-admittedly poor (score 0), in self imposed separation from the current consensus about the prevailing state of the art (which, as before, is scored as 2-3). The outcome of review would correspondingly be given a low score.

The scores imputed here for case 4 merely correspond to the present author's current reading of the field (its models and its community of modellers). In contrast to those for the two examples summarised earlier, none of the scores for example 3 result from explicit research designed to elicit 'expert' assessments from among the retail expenditure flow modelling community as a whole (though such research might be worht pursuing), or of particular individuals (for example, Openshaw, who may judge the current state of the art as 'embryonic', and in much need of

some alternative approach). In this connection, it must be understood that the scoring given can hardly be a final judgement on the state of the field: the history of science contains many examples of former consensus being overtaken by widespread adoption of initially 'rebel' ideas. It is merely given to illustrate the kind of signposting of model quality criteria that is possible via the deployment of suitable assessment frameworks.

6.4 POSTSCRIPT ON THE ILLUSTRATIVE EXAMPLES

In summary, the quality assessment scores assigned in each of the illustrative examples are explicit signals about the quantitative information which they accompany. In all cases, they are signals of caution, or assurance, to anyone involved in the use of related quantitative information, along with explicit indication of the counts on which that information is particularly deficient, or particularly strong. Perhaps more insighful than any resulting string of assessment scores is participation in the proecess of their assignment: the frameworks provide a means of structuring wide ranging reflection and debate about aspects of quantification processes too often dealt with only at intuitive or implicit levels (or else ignored), rather than explicitly. In some cases, the primary rationale is for improved interpretation of 'scientific' evidence (as with example 1); in others for heightening awareness of the meaning and significance of policy indicators (as with example 2); and in others, for pedagogic value (as with example 3). If all quantitative

information were of an unequivocal veracity, there would be no role for such reflective assessments, or parent frameworks for quality coding.

7. CONCLUSIONS

The sort of quality assessment frameworks promoted in this paper might best be viewed as attempts to form practical bridges between, on the one hand, the demands for quantitative information and, on the other, awareness of methodological and philosophical problems in its validity (Harvey 1969, Kuhn 1970, Ravetz 1971, Habermas 1972, Johnston 1983). Their bridging capability rests both on a commonality in language - numerical coding for the quality assessment, matching the numerical form of the information being assessed (*18); and on direct correspondence with key elements of the modern philosophy and sociology of scientific knowledge (technical, methodological, epistemological, functional and practical aspects). There are undoubtedly further levels of development and refinement to the sort of quality assessment frameworks discussed in the present paper, which will come with further promotion of related concepts in a wide range of specific contexts.

Beyond these observations about the foundations of the 'bridges' formed in terms of the frameworks given in this paper, there remain a number of open questions about the construction of their visible structures:

- on the blurring of boundaries between criteria and between modes: the extent to which they are legitimately distinct or more subtly interlocking.

- on the conceptual viability of the idea of normative ordering on individual criteria.

- on how to deal with multiple scores on a given criterion

- on the basis upon which criteria may be assigned - expert

judgement, verification, track record, or whatever.

- on the existence of peer disaggreement about what scores should be given; noting, however, that this is likely to be a 'higher' level of argument than quibbles about the surface values of the quantities under consideration (*19).

- on the possibility of aggregating individual scores to a 'super' index (cf Pasquill weather condition index)

- on the possibility of refining criteria and modes for better tailored frameworks for specific contexts.

- on the possibility of attaching weights to criteria to reflect their deemed relative importance, and consequently developing full multicriteria evaluation tools, for those who are so inclined.

- on the selection and refinement of 'standard' subsets for particular types of context.

- on the need for a phase of broader testing and development of possible frameworks in a wide range of contexts: a typology of contexts and a typology of characteristic frameworks to match.

Such questions can be put on the agenda for future research.

****

FOOTNOTES

(*1) Spatial aggregation of data may create spurious associations, and features reported for areas may not provide a good description of the indiiduals who live there (the ecological fallacy); see, for example, Department of the Environment 1987.

(*2) Harvey (1969) pp319-320 talks of predictive validity, content validity and construct validity in his discussion of related issues.

(*3) Such as multi-attribute utility theory; see for example, Keeney and Raiffa (1976).

(*4) Atkin (1974) proposed that hard data could only be the result of observing strict set membership criteria.

(*5) The frameworks in this paper, in effect, are adaptations and extensions of notations recently developed by Funtowicz and Ravetz for the expression and communication of uncertainties surrounding quantitative technical information. They are adapted and extended to widen their applicability, and out of a desire for greater consistency between representational and evaluative aspects.

(*6) Timmermans (1982) speaks of statistical conclusion validity for this; for completeness, we can note that his categories of 'internal' and 'construct' validity are subsumed under 'theoretical development' below, and 'external' validity under extendability.

(*7) There is a wide range of techniques of sensitivity analysis, and a corresponding literature. It should be noted, though, that none extend beyond sources of uncertainty categorised here as 'technical'.

(*8) Corresponding to 'definitions' in the Funtowicz-Ravetz scheme.

(*9) The Funtowicz-Ravetz category of 'Institutional culture', with modes of 'dialogue', 'accommodation', 'obedience', 'evasion', 'no contact', and 'unknown', is a special case of this criterion.

(*10) What Timmermans (1982) refers to as construct validity is a special case of this.

(*11) Following Funtowicz and Ravetz, we can designate as 'mature' a field whose claim to such a descriptor would be disputed only by 'cranks'; and as 'advanced' a field in which 'rebels' may also be in dispute (rebels having some standing among their peer,s unlike cranks). An 'intermediate' field would be characterised by competing schools.

(*12) Though defining peer communities can itself be problematic in some cases.

(*13) There is possibly here a need for further clarification as between review of process and review of product.

(*14) The modifiable areal unit problem.

(*15) The definitions of 'public' and of 'public interest' are both

problematic.

(*16) Some conventional sensitivity analysis had been undertaken.

(*17) This would depend very much on empirical context.

(*18) Criteria of the sort deployed here may be applicable to non-quantitative knowledge, but the frameworks of which they are a part would probably have less of a place in such cases.

(*19) The scoring must be done by qualified individuals. Discrepancies in scoring may hold interesting interpretations, and may, in the end, not be a problem for the framework, but for the field (eg competing schools).

REFERENCES

R H Atkin (1974) Mathematical structure in human affairs, Heineman

D Black (1984) Investigation of possible increased incidence of cancer in West Cumbria, Report of an independent advisory group, HMSO, London.

B Choi (1987) Space and social theory: a geographical critique and reconstruction, PhD Thesis, School of Geography, University of Leeds, Leeds LS2 9JT.

G Clarke (1986) Retail centre usage and structure: empirical and theoretical explorations, PhD Thesis, School of Geography, University of Leeds, Leeds LS2 9JT.

H Couclelis (1983) On some problems in defining sets for Q-analysis, Environment and Planning B, 10, 423-438.

Department of Energy (1985) 'Get more for your Monergy' Energy Efficiency Office press release 300/9/85.

Department of the Environment (1987) Handling geographic information, Report of the Committee of Enquiry chaired by Lord Chorley, HMSO, London.

S O Funtowicz and J R Ravetz (1986) Policy related research: a notational scheme for the expression of quantitative technical information, Journal of the Operational Research Society, 37, 3, 1-5.

C M Guy (1987) Recent advances in spatial interaction modelling: an application to the forecasting of shopping travel, Environment and Planning A, 19, 173-186.

J Habermas (1972) Knowledge and human interests, London, Heinemann.

D Harvey (1969) Explanation in Geography, London, Arnold.

R J Johnston (1983) Philosophy and human geography, London, Arnold.

A Kaplan (1964) The conduct of inquiry, San Fransisco.

R Keeney and H Raiffa (1976) Decisions with multiple objectives: preferences and value trade-offs, Wiley.

T S Kuhn (1970) The structure of scientific revolutions, University of Chicago Press.

S M Macgill (1983) The Q-controversy: issues and non-issues, Environment and Planning B, 10, 371-380.

S M Macgill and S O Funtowicz (1987) The 'pedigree' of radiation estimates: an exploratory analysis in the context of exposure of young people in Seascale as a result of Sellafield discharges, Working Paper 483, School of Geography, University of Leeds, Leeds LS2 9JT.

S M Macgill and B Sheldrick (1987) Monergy: qualifying imperfect measures of need and of performance, Government and Policy, forthcoming.

J R Ravetz (1971) Scientific knowledge and its social problems, Oxford University Press.

S Openshaw (1987) Computer modelling in human geography, paper presented to the IBG Quantitative Methods Study Group Oxford centenary meeting.

H J P Timmermans (1982) On the validity of mathematical shopping models for estimating turnover of shopping centres, paper presented to the PTRC summer annual meeting, University of Warwick.

A Weinberg (1972) Science and Trans-science, Minerva, 10, 209-222.

A G Wilson (1986) Store and shopping centre location and size - a review of British research and practice, Working Paper 455, School of Geograpy, University of Leeds, Leeds LS2 9JT.

Table 1. CRITERIA FOR QUALITY ASSESSMENT AND DISCLOSURE

CRITERIA \ SCORES	4	3	2	1	0	–
TECHNICAL						
PRECISION	Superlative	Good	Moot	Remis	Reckless	Unknown
ACCURACY	Perfect	High	Medium	Doubtful	Poor	Unknown
ROBUSTNESS	Strong	Resilient	Elastic	Weak	Wild	Unknown
METHODOLOGICAL						
+DEFINITION VALIDITY	Fundamtl/ primary	Standard	Conven- ience	Symbolic	Inertia	Unknown
*DATA COLLECTION	Bespoke/ (task-lab)	Direct survey/fld	Calc'd (ind est)	Educated guess	Uneducated guess	Unknown
+PROCEDURAL COHESION	Full	Strongly linked	Linked	Weak linked	Independent	Unknown
+THEORETICAL DEVELOPMENT	Laws	Well tst'd theories	Emerging theories	Hypotheses	Working defns	Unknown
EPISTEMOLOGICAL						
+STATE OF THE ART	Mature	Advanced	Inter mediate	Embryonic	No opinion	Unknown
EXTENT OF REVIEW	Wide	Fair	Limited	Low	None	Unknown
*OUTCOME OF REVIEW	Total accept	High	Medium	Low	Poor	Unknown

Table 1. (cont'd)

CRITERIA \ SCORES	4	3	2	1	0	-
FUNCTIONAL						
RELEVANCE	Direct	Indirect	Middling/ sufficient	Opport- unist	Spurious	Unknown
CONSEQU- ENCE	Distinc- tion	Merit- orious	Fair	Poor	Negative	Unknown
PURCHASE	Emancip atory	Demo cratic	Mixed	Mild strategic	Private interest	Unknown
PRACTICAL						
EASE OF USE	Fool proof	User freindly	Moderate	User demanding	User hostile	Unknown
COMPUTATIONAL DEMANDS	None	Slight	Moderate	Heavy	Formidable	Unknown
FLEXIBILITY	Great	Wide	Moderate	Narrow	None	Unknown

* = given in original Funtowicz-Ravetz scheme
+ = modified from original Funtowicz-Ravetz scheme

Table 2. QUALITY ASSESSMENT SUMMARY FOR THE THREE ILLUSTRATIVE EXAMPLES

EXAMPLE 1. ESTIMATES OF RADIATION EXPOSURE FROM NUCLEAR INSTALLATION DISCHARGES

CRITERIA	DATA COLLECTION	THEOR DEVELOP'T	STATE OF THE ART	OUTCOME OF REVIEW
Stage 1	1-3	3	2-3	3
Stage 2	1-3	3	2-3	2
Stage 3	1-2	1-3	1-2	1-2

EXAMPLE 2. INDICATORS FOR UK ENERGY EFFICIENCY CAMPAIGNS

CRITERIA	ACCURACY	DEFINI-TIONS	DATA COLLECTION	RELEVANCE	CONSEQUENCE
Indicator 1	1-2	3	0-4	4	1-2
Indicator 2	1-2	2	2-3	2-3	1-2
Indicator 3	1-2	1-2	2-3	2-3	2-3
Indicator 4	1-2	2	1-3	1-3	1-4

EXAMPLE 3. RETAIL EXPENDITURE FLOW FORECASTING

CRITERIA	PRECISION	ACCURACY	DATA COLLECTION	THEOR'L DEV	STATE OF THE ART	OUTCOME OF REVIEW
Case 1	2	1	3	0	2-3	1
Case 2	3	2	2	2-3	2-3	3
Case 3a	2	1	3	0	2-3	1
Case 3b	3	2-3	2	2(3?)	2-3	3
Case 4	3	2-3	3	0	2-3	1