Survey & Statistical Computing

76 Blake Road, Bounds Green, London N11 2AH, UK
Andrew Westlake, MA, MSc, FBCS, CITP, FSS

Phone: +44 (0) 20 8374 4723
Fax: +44 (0) 87 0055 2953
E-Mail: AJW@SaSC.co.uk
Web: WWW.SaSC.co.uk

Combining Data and Knowledge in Models: Promises and Problems

Andrew Westlake

Abstract

We collect data in order to increase our knowledge, but we always have some knowledge before we start. Our existing knowledge raises the questions for which we need more information, and it also guides us in deciding what further data to collect and how to collect it.

Models allow us to generalise from specific observed data to a wider situation. When we analyse data we (usually) update our knowledge. If we can find a formal representation for our knowledge, then a standard statistical technique provides a way to formalise the process of updating our knowledge. This can be the basis for the integration of multiple data sets that relate to different aspects of the same system.

While of general importance, this approach is the **only** way of developing a coherent and integrated understanding of complex systems which are too extensive to observe with a single data set.

But complex methodology is difficult to understand, so we must address the issues of convincing users from the application domain that our models are appropriate and valid, and of making the results obtained from the methodology accessible.

This paper was originally prepared for a keynote presentation at the Association for Survey Computing conference on Maximising Data Value, September 2005.

Keywords

Statistical Models, Knowledge, Uncertainty, Meta-data, Bayesian Methodology, Opus Project, Data Integration

Contents

1. Introduction

1.1 Overview

We collect data in order to increase our knowledge, but we always have some knowledge before we start. Our existing knowledge raises the questions for which we need more information, and it also guides us in deciding what further data to collect and how to collect it. This paper explores this idea, and examines how we can formalise it. By introducing such formalisation into statistical models we can see statistical analysis and the production of statistical results as a process, rather than as a set of independent steps. We can also use this process approach to tackle the problem of coherently combining evidence from different sources which all tell us something about the same underlying system.

1.2 Knowledge and Models

Statistical ***models*** allow us to generalise from specific observed data to a wider situation. Sometimes our models are rather informal, and we think of them as assumptions (which are sometimes not made explicit). A very common assumption is that the process by which we select a sample provides a random, unbiased subset of the population about which we wish to generalise. Similarly, whenever we use a statistical method to compute a significance level or a confidence interval, we are assuming some sort of statistical distribution model. For the standard tests such as a t-test we assume Normality for the statistic. Distribution-free (or non-parametric) tests make weaker assumptions (generally based on independence), and Bootstrap methods again assume that the sample is fully representative of the population. But there always is a model, even if we do not think about it explicitly.

When we analyse data we (usually) update our ***knowledge***. This new knowledge then feeds forwards into our next data collection operation. So it can be useful to think of data collection and analysis not as a single task, but as a step in a continual ***process*** in which knowledge is continually updated as new ***evidence*** is extracted from additional data.

$$\textbf{Model + Knowledge}_i \textbf{ + Evidence}_i \rightarrow \textbf{Model + Knowledge}_{i+1}$$

Many large-scale government business and social surveys are seen as processes, and this view is often appropriate for the analysis of more continuous data capture systems such as retail sales or traffic monitoring. This step-wise approach can also be used where we have multiple sources of evidence (i.e. multiple data sources) which are too complex to be included together in a fitting process – we just fit them in sequence. If we are concerned that the order of fitting has any influence we can iterate until the knowledge stabilises with a balance between the different sources.

How do we formalise this process approach[1] to knowledge updating? The solution is a standard part of statistical theory, which is simple to describe conceptually, if not always easy to implement in practice.

[1] This approach applies to quantitative knowledge, such as about underlying measurements or the probability of discrete options, assessed through evidence in data. Other approaches also have their place. For example, qualitative tools such as focus groups have an important role in the generation of hypotheses, which can then be tested by quantitative means. Similarly, ontological analysis of terms and concepts is invaluable in the structuring of ideas and knowledge about classification structures. We do not discuss these other approaches further in this paper.

1.3 Implementation and Validation

The generic approach gives great freedom for constructing models that cover all facets of our knowledge about a system. However, this generality can cause difficulties in specification, understanding, communication and estimation. Several useful classes of model have been explored, which limit flexibility to some extent, but are easier to specify and explain. Amongst these is the class of Graphical Models (and the related group of Bayesian Networks), which are based on the concept of conditional independence. These are generally easy to conceptualise and explain, and have implementation advantages, though they do have some limitations in terms of the forms of model that are possible. As in many situations, there will be trade-offs to be made in the model specification stage between flexibility and tractability. The effect of these may need to be explored through validation.

In any particular application domain there will be a body of generally accepted theory and knowledge about how aspects of that domain are related and interact. We can use this knowledge to build a generalised *a-priori* model of the domain that can be widely agreed and accepted. This then becomes the starting point for a more specific model of a specific system within the domain. We extend the generic model with the additional knowledge that is specific to the system to be studied.

Domain practitioners often complain that models are ‘black boxes’, and so are not to be trusted. This is not unreasonable, particularly with complex models of the type discussed here, and so must be addressed. We must be able to expose the structure and form of the assumptions made within a model. We must be able to provide information about which datasets were used when fitting a model, and be able to report how much and in what way they influenced the final form of the model. And we must be able to demonstrate the likely validity of the results from the model – generally, this validation will take the form of comparing the predictions from the model with actual data.

2. Models are everywhere

2.1 What sort of model?

The term ***Model*** is very widely used, and can be confusing because it implies different things to different people. Formally, a model is some abstraction (often but not always in mathematical form) representing part of the behaviour of some real-world system, selected in a particular context for a particular purpose. An often quoted remark, attributed to the statistician Prof. James Durbin, is that *all models are wrong, but some models are useful.*

Statistical Models are used to represent the relationships between observable measurements on a real system, in a way that permits estimation of (unobservable) characteristics of the real system (often referred to as parameters). A crucial component of any statistical model is the explicit choice of statistical distributions (with parameters) to represent the variability of measurements.

In computing, the term modelling is used for various processes. Data Models show the structures needed to store the various types of data used in a computer system. These are often based on the Relational or on the Object-Oriented frameworks for data structures. Process models relate to the flow of information between structures and the processes that it goes through.

Conceptual models are ways of organising the ideas (and concepts) used in some domain, together with the relationships and terminology used to refer to them. Most mental models (in our heads) are examples of conceptual models.

In many situations there are modelling frameworks that have been identified to generalise and support the process of constructing and using models. In the statistical field examples are the Generalised Linear Model framework (GLM), and the Bayesian Modelling framework. In database modelling the Relational Database Model (RDBM) plays such a role. In computing, a widely used framework is the Unified Modelling Language ([UML]). This focuses on the production of Object-Oriented computer software, but is also widely applicable for the design of structures and processes. These frameworks are all examples of *meta-models*, that is they are models for the process of producing models.

It is important to recognise that different models of any system can exist with different focus or with different levels of abstraction, and all can be appropriate for their intended purpose. Confusion can arise from failing to recognise the level to which a particular construct contributes, or at which a discussion about a model is taking place.

We have found it useful to explicitly separate out those generalised models which *represent* the general knowledge about the nature of relationships and influences within a domain, in contrast to the specific and detailed models that are used to *explore* our understanding or knowledge about a specific issue or system. Thus a generalised model will make only statements about which measures are related and what the pathways of influence are, whereas a detailed model will need to be specific about the mathematics of relationship and the sources and forms of statistical variability.

2.2 Why Statistical Models?

Statistical models can allow us to generalise from specific observed data to a wider situation.

For example, in a transport system, we may count (in some reliable way) the number of vehicles using a particular segment of road, and we may stop and interview a sample of the travellers and ask what trip they are making (their origin, destination and purpose). This can tell us a great deal about what is happening at the precise location where the measurements are made on the day (or days) of observation, but, of itself, can tell us nothing about any other circumstances. If we want to generalise any of the results from the observations we need to make assumptions or otherwise formulate how our observations might relate to those on a different day, or at a different location. For example, we might assume that the breakdown of trips observed from the interviewed travellers applies to all the travellers who were not observed, and is the same on other days, even if the overall level of traffic changes. Similarly, we might establish relationships between the levels of traffic at different locations, so that measuring levels at one or more locations would allow us to produce estimates of levels at other locations.

So the purpose of creating a statistical model is to allow us to extract evidence from data about some real system in an organised and coherent way. We can then make inferences (or produce estimates) about the real system. A stochastic model will allow us to make inferences about the most likely values and variability of future observations on the system, though we will generally be more interested in estimates of underlying rates and averages. If the model is formulated in an appropriate way we can make inferences about the effect of combinations of factors for which we have no actual data.

For example, while it is at least conceptually possible to collect information (using automated systems) about the number of people entering or leaving every station on the London Underground system every day, it is impossible to conduct interviews to gather information about trip patterns at every one. However, it is perfectly possible to conduct a programme of interviewing which covers all stations over a period of time. A statistical model then allows us to bring all this information together, by making assumptions about the way in which in which demand at different points and on different days is related. We may also make use about other information, such as the connectivity between stations. The statistical nature of such a model is important, because any trip information is based on a sample of travellers, automatic counting systems may have omissions and biases, and because the behaviour of individuals is not constant from day to day.

Sometimes our models are rather informal, and we think of them as assumptions (which are sometimes not made explicit). A very common assumption is that the process by which we select a sample provides a random, unbiased subset of the population about which we wish to generalise. Similarly, whenever we use a statistical method to compute a significance level or a confidence interval, we are assuming some sort of statistical distribution model. For the standard tests such as a t-test we assume Normality for the statistic. Distribution-free (or non-parametric) tests make weaker assumptions (generally based on independence), and Bootstrap methods again assume that the sample is fully representative of the population. But there always is a model, even if we do not think about it explicitly.

2.3 Using Statistical Models

2.3.1 Models vs. Data

Practitioners are often sceptical about results from models, preferring to rely on results derived directly from a particular dataset. While this attitude is understandable, it does ignore the limited applicability of a particular dataset (to what extent can the results be generalised) or the biases inherent in particular data collection methods (what do we do when different datasets give different results).

In practice, all data analysis involves some form of model, even if this is not made explicit. By making the model explicit we are better able to balance information from different sources, understand biases and generalise to the whole system.

2.3.2 Conflicts between datasets

The need for this comes to the fore when we have different datasets that give different answers (estimates) for the same question. For example, the UK census asked people about the location of their main work, and, from their known home location, was able to compute information about demand for travel to work between various origin and destination zones. The London Area Transport Survey (LATS) conducted household interviews at around the same time in which respondents kept a diary of their travel behaviour, from which similar origin-destination demand patterns for work were derived. The results differ substantially, often by 30%. Similarly, LATS roadside interviews ask about origin and destination of the current trip, and the demand estimates from this data are again different.

The practitioner’s response to these differences is usually to ask which is right and can be trusted, and which is wrong and should be discarded.

The statistician’s response is to say that (probably) they are all correct but they are just different. The source of this difference is rarely just statistical variability – in the LATS case all the sample sizes are easily big enough to be reliable. Rather the differences are due to biases (systematic differences) of some sort. Generally there are two reasons for such bias.

1. The questions are different. This is clearly the case with the census and LATS household data. The census asks about ‘usual’ place of work, whereas LATS records actual travel to work, which will not always be to the usual place.

2. The sample selection processes are different, so the different subgroups of the population (with different behaviour) are present in different proportions. This applies to the LATS household and roadside data, in several ways. The household survey has a rigorously defined sample selection process, but the drop-out process is biased with regard to household size. In contrast the selection process for the roadside interviews is based on randomly selecting and stopping passing vehicles, which has a different drop-out process. Similarly, whereas the household diary covers all trips, only a subset of trips can be detected at the roadside.

To obtain coherent information about a system from multiple data sources we must take into account the differences in the processes that yield the different datasets, and we do this by introducing these as factors in our statistical model.

2.3.3 Using results from model

How do we persuade practitioners that models (and the results from them) are valid and useful? We address this through the use of meta-data about the models and the model fitting processes. From a philosophical perspective we would argue that there is always a model, so it is better to understand it and be able to criticise it than to pretend that there is no model. However, rather than trying to win a philosophical argument we concentrate on exposing the qualities of a model so that users can make judgements as to the usefulness of model results. We focus on providing information to users about the provenance and reliability of results obtained from a model.

3. Types of knowledge

The knowledge that we have about a system can take many forms.

3.1.1 Knowledge about the System

When we decide to collect data about some system or process it is often because we recognise a gap or deficiency in our existing knowledge. With continuous data collection it is because the system is continually evolving and we want to update our knowledge about its current state.

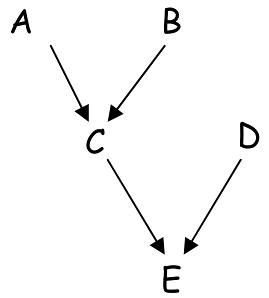
When we design a survey we use our existing knowledge to decide which questions to ask, how to formulate them and how to bring them together (in a questionnaire, perhaps) in a way that is likely to yield of high quality, i.e. that accurately reflects the state of the system (or person) being questioned. If we have a choice about data collection methods then we use our knowledge to assess which will be the most appropriate (usually most cost effective) single or combination of methods to use. Finally, having identified the target population about which we want to collect information, we design a sampling process for the selection of units (usually people) about whom data will be collected. This latter is often the only factor that is explicitly associated with the collected data.

3.1.2 Knowledge about Values

We may know something about the range and distribution of values associated with variables: *from past data we have estimated the mean value of C to be m (with standard error s_m) and the standard deviation to be s_p. The distribution looks approximately Normal, but there is a suggestion of skewness with a long upper tail.*

We use this type of knowledge in designing data collection (for example in power calculations to determine sample size), and it can influence the choice of analysis techniques and hypotheses, but with classical approaches to analysis there is no way to use it directly within the analysis itself. Bayesian methods, on the other hand, explicitly allow such prior knowledge to be included in the analysis process.

3.1.3 Knowledge about Relationships



We may have ideas about the way in which factors interact and influence each other: *A and B both influence C, which in turn, with D, influences E.* We may have ideas about the form of relationship between factors: *E increases as D increases*.

We may have knowledge about the value of parameters in relationships. *We think the slope of the relationship between D and E is about 1.5, certainly near D=100* or *we think that about 30% of the variability in C can be explained by A*.

We use this knowledge to make choices about the appropriate form for the statistical analysis that we apply to the collected data. For example, using regression analysis to explore the effect of A and B on C implies the assumption that both affect C linearly and that their effects are additive.

Notice that we can include ideas about relationships (both form and context) and about distributions. The latter allows us to include uncertainty in the model. This can be related to variability in the observation process (whether inherent to measurement or from sampling) and to the detailed form of the model. We can also express uncertainty about the exact form of components of the model, generally by using wider classes of relationship or distribution and including uncertainty about the parameters that determine the specific form. These ideas are expanded later.

4. Representation of knowledge and uncertainty

4.1 From Confidence to Uncertainty

In simple statistical analysis we represent the uncertainty associated with an estimate of a parameter by calculating a confidence interval. For different levels of confidence we obtain different intervals (or limits) and we can represent the set all limits as a distribution over the possible parameter values. In many cases this will take the shape of a Normal distribution, because the Normal distribution is assumed for the data.

4.1.1 Confidence Limits

If we have an estimate $\bar{x}$ of the mean μ, then we usually calculate confidence limits for the true value of μ in the form $F_{\bar{x}}^{-1}(\alpha)$, $F_{\bar{x}}^{-1}(1-\alpha)$, where $F_{\bar{x}}^{-1}(\alpha)$ is the inverse cumulative distribution function[2] (CDF) for $\bar{x}$. This yields a confidence interval of size $1-2\alpha$. The size is the probability that the interval (which is a random variable because it is calculated from the data) actually includes the true value of μ.

[2] Strictly speaking, the confidence limits can also be dependent on the true value of the mean, but we ignore that detail in this discussion.

If we assume Normality for the distribution of $\overline{x}$ and flesh out the context to have n observations from which we calculate an estimate $\overline{x}$ plus a variance estimate s^2, the limits simplify to $\overline{x} \pm \frac{s}{\sqrt{n}} \Phi^{-1}(\alpha)$, where $\Phi^{-1}(\alpha)$ is the inverse of the cumulative distribution function of the standard Normal distribution (zero mean and unit variance).

4.1.2 Confidence Curves

We usually work with a single confidence interval for a given parameter. However, the size to be used for this is a matter of judgement (or just arbitrarily chosen), and in fact there is a continuous range of possible confidence intervals that could be used. An alternative approach is to represent them all.

As a concrete example, if $\overline{x}$=50, s^2=64 and n=100, the (Normal) cumulative function is as shown in the diagram on the right. The horizontal (x) axis corresponds to possible values of the true mean, and the vertical (y) axis shows the probability that the true value is lower than the corresponding x value.

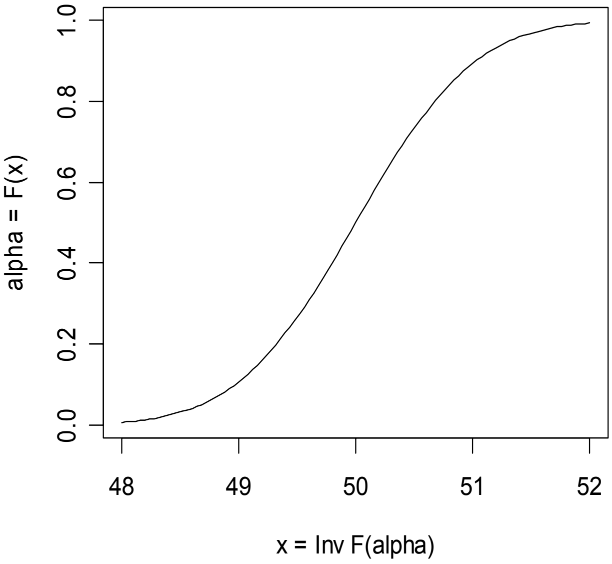


Figure 1 CDF of an Estimate as a Confidence Curve

This diagram can be used to obtain a confidence interval of any size – just take the appropriate probability points on the y-axis and read off the corresponding confidence limits on the x-axis. This example uses the Normal function, but any CDF can be used, as appropriate for the data.

4.1.3 Confidence Distribution

The same information can be shown in the form of the corresponding probability density function (PDF), because the CDF is the integral of the PDF. In this diagram, the confidence associated with any possible range of values for the true mean is the area under the graph between those values. In our example the density function is for a Normal distribution, but, again, it could be any distribution that was appropriate to the context.

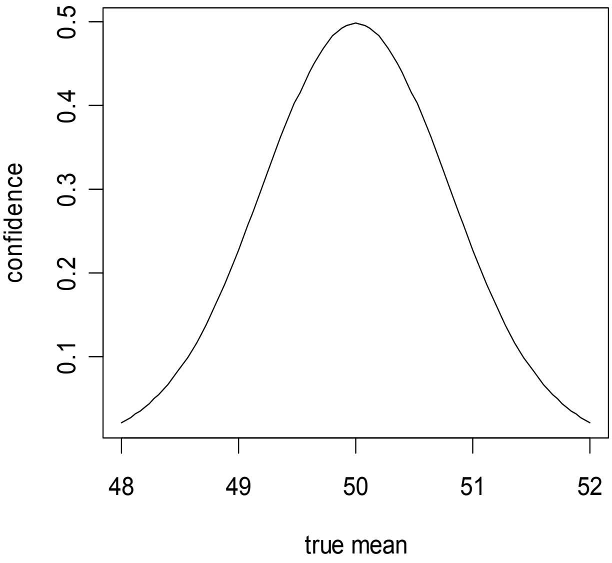


Figure 2 PDF of an Estimate as a Confidence Distribution

The point of this is to show that the information about the true value of a parameter that we use in deriving a confidence interval can also be represented as a distribution (PDF). The cumulative form is easier to work with if we do want to compute confidence limits, but the density form can be used in other ways, as we shall see.

4.2 Uncertainty Distributions

4.2.1 Knowledge as Uncertainty Distribution

So far we have used the classical terminology of confidence to talk about our knowledge of the range of possible true values for the parameter. We can equally talk about our uncertainty about the true value. Then the density function above can be thought of as an uncertainty distribution – a flatter distribution corresponds to greater uncertainty, and a more peaked one to lower uncertainty, or more precise knowledge.

Representing our uncertainty about a parameter as a distribution does not imply that the parameter is a random variable. The parameter is a fixed property of the reality about which we have collected data, and it is our uncertainty that is represented by the distribution. So we can use the mathematical properties of these functions and manipulate them as we do probabilities, but we must remember that their interpretation is not as probabilities.

We can take the idea further, and represent any uncertainty with a distribution. So far we have talked about representing the uncertainty that remains after extracting knowledge from a particular dataset by performing a standard statistical calculation. But we have knowledge (and limits on uncertainty) from other sources as well. Why not also represent this uncertainty in the form of distributions?

We do not require that the uncertainty distribution is derived from data, we can simply 'invent' it. Of course, it is not sensible to do this without some prior knowledge, or justification, to support the particular choices that we make.

For example, we could express our knowledge about a regression relationship between x and y as:

$y \sim \mathcal{N}\left(\alpha + \beta \times (x - 100), \tau\right)$, where τ is a precision[3] (or inverse variance) parameter,

$\beta \sim \mathcal{N}(1.5,1)$ – we are reasonably confident about a value near 1.5 for the slope,

$\alpha \sim \mathcal{N}(200,.001)$ – we think y is around 200 when x is 100, but are not at all confident about this,

$\tau \sim \mathcal{T}(.001,.001)$ – we know very little about the variability[4] of y around the regression line.

The first equation is a statement about the form of model and variability (Normal) for the real system, whereas the following three are statements about uncertainty over the parameters of the model.

5. Formulating Statistical Models

5.1 Model Structure

The heart of a model is a specification in mathematical terms (i.e. largely algebra) of the factors (variables) that influence the measures of interest in the system being studied, and the way in which

[3] This formulation (with $\tau = \frac{1}{\sigma^2}$) is often used for uncertainty, with the justification that low precision (τ near 0) is a more natural representation of great uncertainty than is a large value of the variance σ^2.

[4] The Gamma distribution (which is always positive) is often used to represent uncertainty about precision parameters.

they interact in their influence. Of course, the particular factors and form of relationships will be specific to the problem we are addressing. Our focus is on extracting coherent knowledge about the underlying system from the available data.

5.1.1 Variables and Relationships

Variables relate to the data subjects (or various types), and their values are not generally of direct interest in themselves. We are interested in what they tell us about the underlying system (or population of data subjects).

The variables will have statistical distributions associated with them to represent variability (i.e. they are not necessarily assumed to be fixed). This variability might in part be due to measurement error, where the data recorded does not correspond exactly to the value being measured. This can happen, for example, when a vehicle counting device does not accurately register every individual vehicle, or when a respondent is uncertain about the overall income for their household. It is important not to confuse measurement error with bias in the measurement process – the latter should be included explicitly in the mathematical part of the model. Sampling issues may also be important – we return to this later.

Variability also represents unpredictability of actions and events. For example, people with the same set of background characteristics, including their jobs and home location, will make different decisions about their transport needs, in ways that are not predictable without very much greater depth of understanding (and modelling) than is feasible. And the same person may make different decisions about travel on different occasions. Also, the time taken for a particular journey may have unpredictable variations as a result of bad parking or minor traffic accidents or road works. There will also be larger effects on the travel time, consequential on the weather or the traffic loading, which we may choose to include explicitly in the model, but effects below the level of interest for the model will be treated as unpredictable variability.

We will need to be explicit about the location and nature of variability in the model, including assigning specific distributional forms (often we will assume the Normal distribution). The parameters of these distributions are included with the measures that we are interested in estimating.

Relationships can be derivations, showing how one variable is derived from others, or they can be stochastic, saying that the parameters of the distribution of a variable depend on functions of other variables (and parameters).

It is sometimes appropriate to interpret relationships as constraints. For example, a derivation equation can be thought of as an equality constraint, and a distribution where the mean is a function of other variables (as in regression) is a distributional constraint, giving the likelihood of a particular range of values.

5.1.2 Parameters

Conceptually, parameters relate to the true characteristics of the underlying system, as viewed through the model – they are the things about which we are trying to extract and update our knowledge. For example, in the transport context we will have parameters that relate to the probability (or rate) that people with particular demographic characteristics, living in a particular area, will want to make a particular journey for a particular purpose.

We use parameters in relationships between variables, and in the distributions we associate with variables. Familiar examples are the slope and intercept in a regression model, and the mean and variance of a Normal distribution.

Similarly, we can have relationships between parameters, and these relationships can have further parameters. Also, parameters can be defined in terms of statistical distributions, which again have parameters. Parameters used to determine other parameters are sometimes referred to as *hyper*parameters. Depending on what is appropriate for the formulation (or parameterisation) of the model, these may not correspond to measures of direct interest, but (through the relationships) they can be used to derive those of direct interest.

Generally we will have some prior knowledge or uncertainty about the parameters, which will be more or less informative depending on what experience we can bring to the context and the understanding of the model.

The distinction between variables and parameters sometimes is a little fuzzy, particularly where the real system of interest includes components and data at different levels. So while this is a useful construct for thinking about models, we do not try to impose the distinction where it is not obvious.

5.1.3 Model Uncertainty

As well as uncertainty about the values of parameters in the model, we may be uncertain about the appropriate form for the model. We can cope with this by introducing parameters to control the functional form of the model, in addition to those that relate directly to the underlying system.

For example, to generalise our previous regression example of uncertainty, if we believe that a regression line may not be straight, but could be a monotonic concave or convex curve, we could use the Box-Cox transformation. This replaces x with x^{λ}, where λ is a curvature parameter – $\lambda > 1$ curves upwards, so y increases faster for larger x values, while with $\lambda < 1$ the slope declines (while remaining positive). Figure 3 shows examples of this curvature, where the precise parameterisation has been chosen to retain the slope of 1.5 when x = 100.

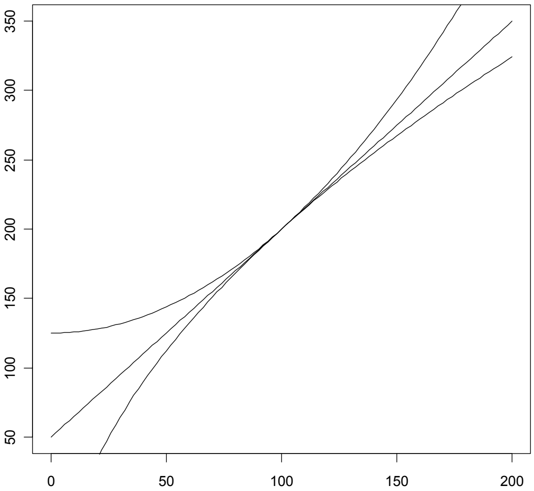


Figure 3 Box-Cox function with λ = 0.5, 1 and 2

This simple example shows how we can convert uncertainty about the form of a relationship into uncertainty about a parameter, by introducing a more general (parameterised) form of the relationship, in which our best guess is a special case. A similar approach can be used when we are uncertain about the appropriateness of particular distributional forms, since all commonly used distributions are special cases of more general forms.

5.2 Bayesian Approach

In simple statistical analysis we represent the uncertainty associated with an estimate of a parameter by calculating a confidence interval. For different levels of confidence we obtain different intervals (or limits) and we can represent the set all limits as a distribution over the possible parameter values. In many cases this will take the shape of a Normal distribution, because the Normal distribution is assumed for the data.

Although we can represent our uncertainty about a parameter as a distribution, this does not mean that the parameter is a random variable. Rather, it is a fixed property of the reality about which we have collected data, and it is our uncertainty that is represented by the distribution.

We can take the idea further, and represent any uncertainty with a distribution. Thus we do not require that the distribution is derived from data, we can simply invent it. Of course, it is not sensible to do this without some prior knowledge, or justification, to support the particular choices that we make. Where we do have knowledge about the parameter we tend to talk about knowledge rather them uncertainty distributions.

With uncertainty represented in the form of distributions, we can draw on what is known as Bayesian Methodology for working with our models.

Bayes' theorem is a simple statement about conditional probability. It comes from the recognition that a joint probability can be written as the product of conditional and marginal probabilities.

> $\mathrm{P}(\mathrm{A} \wedge \mathrm{B}) = \mathrm{P}(\mathrm{A}|\mathrm{B}) \times \mathrm{P}(\mathrm{B})$ – that is, the probability of both events A and B occurring at the same time can be calculated as the probability of B multiplied by the probability of A given that B has already occurred.

Bayes' original use of this was to show how to calculate conditional probabilities, as:

$$\mathrm{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P}(\mathrm{A} \wedge \mathrm{B})}{\mathrm{P}(\mathrm{B})}$$

In contrast, Bayesian Methodology uses the first formula twice to show how to reverse the ordering of the conditioning.

$$\mathrm{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P}(\mathrm{B}|\mathrm{A}) \times \mathrm{P}(\mathrm{A})}{\mathrm{P}(\mathrm{B})}$$

If we now substitute a parameter θ for A and consider B to be our data X, we have:

$$\mathrm{P}(\theta|\mathrm{X}) = \frac{\mathrm{P}(\mathrm{X}|\theta) \times \mathrm{P}(\theta)}{\mathrm{P}(\mathrm{X})}$$

This formula says that where we have prior knowledge about the parameter θ in the form of an uncertainty distribution $\mathrm{P}(\theta)$, we can update this knowledge if we also know how the data distribution $\mathrm{P}(\mathrm{X}|\theta)$ depends on the parameter, obtaining the posterior knowledge distribution $\mathrm{P}(\theta|\mathrm{X})$. Note that we do not need to evaluate $\mathrm{P}(\mathrm{X})$, because it does not depend on θ, so acts as a normalising constant, and we can normalise the posterior distribution directly.

With our statistical model, if we use Bayesian methods we can say that the model gives us $\mathrm{P}(\mathrm{X}|\theta)$ and so tells us how to combine the evidence from the data with our prior knowledge $\mathrm{P}(\theta)$, resulting in improved (posterior) knowledge $\mathrm{P}(\theta|\mathrm{X})$. Of course, θ will be a complex structure of parameters, and we probably work in terms of density functions rather than discrete probabilities.

5.3 Updating knowledge

Given a complete model (including out prior distributional knowledge) we can predict what the results might be when we observe a particular part of the underlying system. If we have real observations on this part of the system, we can compare the characteristics of the observed data (the evidence) with these predictions. Using evidence extracted from the data we can then update (or **calibrate**) the distributional parts of the model to bring the predictions closer to the evidence. In this way our knowledge is updated, and we can talk about the posterior knowledge after incorporating the evidence. Unlike the classical approach to model fitting which uses only evidence in the current dataset, the Bayesian approach balances the new evidence with the knowledge already in the prior distributions.

Although the Bayes updating formula can in theory be evaluated explicitly, with most complex models this is intractable. This is where the MCMC approach is used – the names Metropolis-Hastings and Gibbs Sampler are also used in this context. In effect this approach randomly generates a sequence of observations from the prior distribution of the parameter. At each step it uses the likelihood of the data under the previous and proposed parameter values in a rule that determines whether to accept the new parameter value or not. At the end of a large number of repetitions of this process we have an empirical distribution of the parameter, conditional on the data, which is an estimate of the required posterior distribution.

A number of important points should be noted

1. The model (the mathematical relationships and distributions used for calculating $\mathrm{P}\left(\mathrm{X}\middle|\theta\right)$) is not changed by the updating process, but the knowledge (which can include information about parameters that control the precise form of the relationships) is changed.
2. The updating (or fitting) process does not produce estimates from the data. Rather it updates the estimates that we can produce from the model by changing our knowledge about the parameters.
3. There is no requirement that any particular data set should contain observations on all the factors contained in the model. If a particular dataset contains no information about a parameter, then the posterior distribution of the parameter will be the same as its prior distribution, i.e. the knowledge (or uncertainty) is unchanged (because $\mathrm{P}\left(\mathrm{X}\middle|\theta\right)=\mathrm{P}\left(\mathrm{X}\right)$). We can also have factors in the model which are inherently unobservable (often called latent factors), but which influence things that can be observed.

5.4 Multiple Sources of Information

Alternative sources of information about any system often produce conflicting estimates of measures of interest. Invariably this is because the sources use different techniques to elicit and gather the information, resulting in differential biases in the data. The sources of these biases may be different survey instruments, different sample selection processes or different standards in the execution of the data collection. This presents no fundamental problem to the model-based approach.

Because all datasets relate to the same underlying system, we only need one model. But we now use multiple fitting steps to bring in the evidence from multiple datasets.

Consider first the situation where we have two independent sets of information about the same (parts of the) system, so providing information about the same (set of) parameters. Independence means that neither dataset depends on the other for a given set of parameter values, so the joint probability can be factorised into two independent parts. We can thus apply the Bayesian approach to the two datasets together.

$$\begin{aligned} P(\theta \mid X_1 \wedge X_2) &\propto P(X_1 \wedge X_2 \mid \theta) \times P(\theta) \\ &= P(X_1 \mid \theta) \times P(X_2 \mid \theta) \times P(\theta) \\ &\propto P(X_1 \mid \theta) \times P(\theta \mid X_2) \\ &\propto P(X_2 \mid \theta) \times P(\theta \mid X_1) \end{aligned}$$

This says that the posterior knowledge extracted from both sets of data can be obtained by using two fitting steps. In the first step we use one dataset to obtain a posterior distribution (from the prior knowledge), and then we use this as the prior knowledge for a second step using the other dataset. The datasets can be used in either order, and the results should be the same (to within the precision of the fitting process). This idea generalises to more than two data sets.

The assumption of independence of the datasets is reasonable for independent samples or surveys that are about the same aspects of the real system. Where there are differences in the sampling or data collection methods, for example, between household and roadside interviews, the model (i.e. the $\mathrm{P}(\mathrm{X}|\theta)$) must accommodate this.

The assumption of independence may not be reasonable for all situations, and then the order of the fitting steps may make a difference. In that case we need to iterate the fitting process, using the datasets again, each time starting with the knowledge already extracted, until the posterior knowledge reaches a stable balance point.

6. The Role of Meta-Data

6.1 Meta-data as Audit Trail

Over recent years the concept of meta-data (and the recognition of its importance) has become widespread in many fields. However, the general idea of meta-data has many different applications in different areas and so means different things different people. For example, the Dublin Core proposals (and extensions such as the UK government e-GMS standard) have proved important in the context of resource discovery, especially on the Internet. Related to this is the ISO 11179 standard for meta-data repositories. Similarly, the DDI (Data Documentation Initiative) Codebook standard for the description of survey datasets has achieved wide acceptance. Several examples of the application of this approach to travel survey data are discussed by Levinson and Zofka [LeZo04]. Other authors (for example, Papageorgiou and colleagues [PPTV01]) have made proposals to extend the statistical meta-data concept to give much more complete coverage of statistical data, including sample design and tabulation.

An alternative thread that has received attention in the statistical domain is that of process meta-data. This is information that describes and documents the processes through which data has passed. This can be seen as providing an ***audit trail*** so that it becomes possible to discover details about any transformations, adjustments or corrections that have been made to data before it reaches the form in which is published. This approach to statistical meta-data is discussed by Green and Kent [GrKe02] in one of the deliverables from the MetaNet project [MetaNet].

Also from that project, Froeschl and colleagues [FGdV03] make valuable contributions about the concepts underlying statistical meta-data. Amongst their insights is the useful distinction between what they call Intentional and Extensional meta-data.

Intentional meta-data documents concepts, objectives, reasons and other factors that precede or are external to statistical data. This can include things like decisions about the sample design and data collection methods, the names and coding of variables, and the people, organisations and context associated with data. It is generally textual, and, while the structure of the components will have a formal organisation, the content will be less formally controlled.

Extensional meta-data documents actions and specifications. It includes things such as sample selection rules, derivations and transformations, file locations, process and analysis specifications. It can usually be captured by software processes, and can be part of the input specifications for other processes. The content of such items will have a tight formal specification.

6.2 Meta-data in Opus

The Opus project (funded under the ERPOS component of the European 5th framework) is developing a methodology for the integration of multiple data sources about complex systems. As part of this project we are implementing a system for reporting on the qualitative aspects of the results from the statistical methodology, as an adjunct to the results (estimates) themselves. We do this through the use of meta-data about the statistical models used.

In Opus we focus on process metadata, mostly of extensional form, to support meta-data that contains the specification of the statistical model. Our objective is to keep track of the processes that are applied in developing the statistical model from which conclusions are drawn. We assume the existence of a suitable meta-data model that covers all other aspects of the data.

Details about the Meta-data system adopted by the project appear in the following section, but the main elements are as follows:

- the mathematical specification of the model that is chosen, including all its statistical components
- the model fitting processes that are applied to the model, including all the datasets that are used
- the state of knowledge about the modelled system that is extracted from the data by the fitting processes
- specifications for the results that are extracted or reported from the final model

The intention is to capture all pertinent information about the model fitting process and link this to any results produced from the model. With this information we open up the black box of the model, so that a user can explore the qualities and reasonableness of the model and the fitting processes, and can ask questions about the reliability of results obtained from the model. Because the information is formally structured, it is also possible for other software to read the specifications and use them to repeat the model fitting process (for validation of fitting algorithms), or to apply the same model to different data.

However, while the capture of this information is essential, its mere existence is not sufficient. Facilities are needed to present the information in ways that are accessible to particular groups of user, together with guidance about the types of question that should be asked about the model and the results. This is the objective of the Reliability and Provenance concepts presented later.

7. Representation of Statistical Models and Processes as Meta-Data

7.1 Structures for Meta-data

In the Opus project we use UML to hold specifications of the structures and functionality that we have designed for handling meta-data. Figure 4 shows some of the high level structures that are relevant for this paper. The full structural model contains much more detail. Documents describing the details are available to people who sign up to join the project discussion groups, and will be widely published at the end of the project.

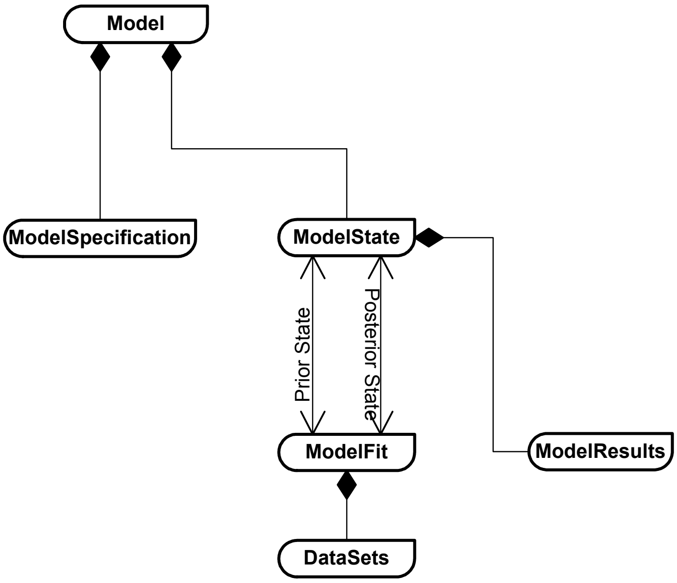


Figure 4 Outline of the Opus Meta-data Structure

The **ModelSpecification** is a (single) complex structure that contains all the information about the form of the model that has been chosen as an appropriate abstraction of the real system. This includes the variables (or factors) about the underlying real system that are pertinent to this model, the parameters that have been chosen to summarise or represent influence mechanisms in the real system, the nature and forms of mathematical and statistical relationships between the variables and the parameters, and the statistical distributions that represent the variability in observations on the system. A considerable amount of structural knowledge and expertise goes into the construction of this specification, and the specification as a whole represents the set of assumptions about the real system that are embodied in the model. The stored meta-data is mostly of extensional form, being formal specifications that can be transformed for use in suitable software, but there is also intentional meta-data that documents reasons for particular model formulations or parameterisations and for making particular assumptions.

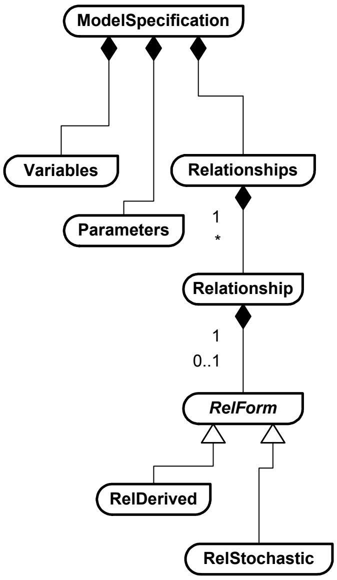


Figure 5 Meta-data for Model Specification

The **ModelState** element represents knowledge about the values of parameters in the model, expressed as uncertainty distributions. Every time we use data to update (or improve) the fit of the model the knowledge changes, so in general we will have a set of states associated with the model.

The **ModelFit** represents the process of using one or more **datasets** in some well-defined methodology to update the knowledge about the system through the model. Such an updating process will start from some state of knowledge about the model (the prior state) and will produce a new state (the posterior state) in which the knowledge (uncertainty distributions) has been updated. Often the overall process of fitting a model will involve a sequence of different fitting steps, in which different datasets are used, perhaps with different fitting methods. Iterative procedures are also possible, in which the model is repeatedly updated from various datasets until stability is reached in the uncertainty distributions. These processes produce chains of model states which represent the fitting sequence.

A fitting step may require mapping between the form of variables in the data and that in the model. For example, the model may be expressed in terms of the behaviour of individuals, but some data might only be available after aggregation. Or individual income may be represented as exact amounts in the model but only collected as banded groups in a survey. There is no problem about this, as long as it is possible to calculate the likelihood of the data that is implied by the model. In practice this means that any link between the model and data that involves variability or uncertainty needs to be represented explicitly in the model, while anything involving deterministic transformations or aggregation can be handled as a data mapping as part of the fitting step.

We assume that individual datasets are accompanied by their own meta-data describing their contents and their collection processes. In the Opus tests we will be using the DDI Codebook for this information.

ModelResults, whether conclusions, estimates or simulated data, are always based on a single state of the model, generally what might be characterised as the ‘final’ state after extracting all available information from all datasets. Generally speaking, results will be obtained by taking the final state of knowledge about one or more parameters and working through the mathematics of the model to be able to make statements about the implications of the model for the underlying system.

7.2 Using Meta-data with Results from a Model

Model results can always be linked back to a single state of the model, from which we have access to both the specification of the model and the chain of fitting steps that led to the final state. Thus software that is designed to support use of results from the model has access to all the meta-data that documents the final state of the model and how this was reached. This is the basis of our efforts to provide users of model results with supporting information about the provenance and reliability of the results. These efforts are discussed below.

8. Results from Statistical Models

8.1 The form of Results from Models

The end result from application of this methodology is a calibrated statistical model. This is specified in terms of a set of mathematical relationships among the variables and parameters of the model, including components that describe the stochastic variability exhibited by the underlying system. In addition, the knowledge about the model parameters that has been extracted from the evidence available in datasets is summarised in terms of posterior distributions which encapsulate the best estimates and our uncertainty about the parameters.

An experienced analyst, familiar with the methodology, can use the model to extract information about the underlying system, covering estimates of measures of interest, their variability, and the uncertainty associated with these estimates. Where dealing directly with the mathematics of the model is too difficult, the implications of the model can be presented in the form of simulated datasets generated from the mathematical specification. A simulated dataset will generally include variability associated with the underlying system, and can also include variability arising from uncertainty about parameter values.

8.2 Provenance and Reliability of Results from Models

We anticipate the presentation of three forms of information derived from a model.

1. **Conclusions**. Summary reports which provide interpretations of the fitted model, based on the experience and judgement of the author. These will be largely textual, but will include illustrative material and links back to the model.

2. **Estimates**. Presentation of the posterior distributions of quantities of interest from the underlying system. This can be done in terms of summary statistics (particularly means and standard deviations) of the posterior distributions, or of complete distributions, presented as histograms or multivariate contour plots (for example). Note that the distribution represents our uncertainty about the true value of the quantity, so it is important to present this as well as any point (best) estimates.
   Population parameters of direct interest to users (for example, in decision making) will be the primary focus, but these are generally dependent on internal (hyper-) parameters, which are the ones directly adjusted by the fitting process. But estimates can be obtained for any derivable measure on the underlying system, with a corresponding derived posterior distribution.

3. **Synthetic data**. Given the model specification and the posterior distributions, it is possible to simulate observations on data subjects. In this way, we can create synthetic datasets which have the same characteristics as the model. These are much easier to analyse for people used to handling real datasets. It is also possible to generate data for specific conditions, for example by limiting the impact of abnormal events, focussing on particular subsets of the overall possibilities, or assuming away some uncertainty in parameters.
   The problem is that synthetic data is not real, and its statistical properties are not the same as those of real observations on the underlying system, because they come entirely from the fitted model. The challenge is to guide users to appreciate these differences.

These three types of information have close parallels with information obtained by more traditional methods. The difference is in the central role of the model in our methodology. Instead of presenting information that is directly derived from a dataset, and which is then inferred to be directly about the underlying system, all our information is mediated by the model. The model serves to balance and explain differences in the results obtained from separate datasets, by requiring that differences in the data collection methods or the response processes are made explicit. It also makes it possible to explore the implications of the model for combinations of circumstances for which no data has actually been observed.

For such results from a model to be useful and usable, the user must have confidence in the model. We must be able to explore and ask questions about the nature and qualities of any fitted model. We thus propose that two additional types of information should be available with all results that are derived from a statistical model.

4. **Provenance**. Information about the structure and objectives of the model (including its mathematical form), and about the model fitting process (the audit trail). This includes information about the fitting methodology (which will apply across a set of related models), together with the datasets used at the various fitting stages and the contribution of each such stage to the final fit. The latter is particularly important in terms of understanding how well the posterior distributions of parameters have been determined by the fitting process.
5. **Reliability**. This relates to the posterior distributions of the model parameters. But instead of focussing on estimates of quantities of interest in the underlying system, it focuses on the uncertainty that remains about the model parameters. We explore whether the parameters are well-determined, the source of the knowledge about a parameter (ie prior knowledge or particular datasets), and how well the final model reproduces the datasets used. It is important to distinguish between *uncertainty* about parameters (which should generally decrease as more data is used or as the model formulation is improved) and *variability* in observed data that is associated with measurement processes or unpredictable behaviour.

The source of most of this information is the meta-data that describes a statistical model and that records (like an audit trail) the processes used to arrive at the final state of the model. We have proposed a structure for meta-data about statistical models that includes (potentially) all this information (it is in effect a complete audit trail for all the specifications and stages used to produce results). But we also need to find ways of presenting this additional information that are accessible and comprehensible for different groups of user.

8.3 Communicating and using knowledge

In a simple situation with a single dataset, the prior knowledge is embedded in the design of the data collection and in the head of the analyst. The evidence from the data is used to update the analyst's head, and some of this updated knowledge gets written down in reports and made available to others. This posterior knowledge may be in the form of estimates of various statistics (such as rates, means and standard deviations), or in the parameters of simple (and easily interpreted) relationships estimated from the data, such as regression slopes. What is appropriate depends on the level of skill and understanding of the analyst, and of the intended target for the knowledge.

This does not work for large and complex systems. The model is complex and covers many situations, while a dataset will refer only to a part of the system. Analysts (users of the data and knowledge) are many, with different requirements, different focuses, and different levels of understanding of the system and the model. Simply providing access to the formal specification of the model and the updated, formally expressed knowledge about the model, will not provide understanding to most domain specialists used to analysing individual datasets, so we need to find other means.

As with other complex systems (such as databases) we need to be able to provide **views** of the system (or the knowledge about it) which address the needs of specific users. To some extent we can do this by re-formulating the mathematics of the model to focus on a specific application. When doing this we can choose to leave out components of the model that are not needed in the application. This can be done by **Conditioning** (setting parameters to fixed values) or by **Marginalising** (averaging over the variability associated with the omitted components).

However, this algebraic manipulation will often not be possible with complex models, and then we have to resort to numerical methods. We can estimate averages for some output measure under differing input assumptions (ie varying input parameters) and display the results in diagrams. This is appropriate if we are wearing our analyst's hat and wish to communicate a particular message to a particular audience. More generally, we can generate synthetic datasets under various assumptions (conditioning and marginalisation, again), which we then pass over for analysis of a more traditional kind.

What can be found from such synthesised data?

Clearly, there can be no knowledge in the synthesised data that is not already in the model, because the data is generated from the model. But the domain analyst can extract knowledge from the synthesised data without needing to understand the complexity of the full model, and without having to deal with the noise in the data that would have come from the components that have been assumed out. It is clearly important to know what the assumptions have been made and to have a good general understanding of the form and limitations of the model, but this is also true (though simpler) with simple observed data sets.

In addition, since we are working from the model and are not constrained to synthesise data that corresponds to real observations, we have a number of forms of additional flexibility.

1. We can produce synthetic data that relates to combinations of factors for which we have no observations. So we could synthesise data about the flow along a traffic link at a time of year when we have no actual observations. We will have observations about the flow on the link at other times, and we have information from other places about how the flows vary over the year.
2. We can produce synthetic data for situations that do not currently exist. For example we could set factors in the model to represent the construction of a new housing development, and then produce data to investigate the impact on existing traffic flows.
3. We are not restricted to observable variables, so synthesised data can include realisations of latent (unobservable) variables.

9. Provenance and Reliability of Model Results

9.1 Objectives

Users of results from statistical models should properly be asking questions about how the results were obtained and how much confidence they should have in conclusions drawn from them. We use the term *Provenance and Reliability* to refer to this area. This covers all issues to do with the understanding and interpretation of fitted models.

Different types of user will expect answers of different complexity and detail. Some answers can be generic, describing the philosophy behind the statistical methodology and Bayesian modelling, or showing the outline of the model fitting processes (perhaps through the use of UML diagrams). Other answers will need to be based on the specific components used in the model from which the data are synthesised, and further ones will make use of the detailed posterior information about the parameters. All this information will be available in the form of metadata, the top-level structure of which has been described above. The same information may need to be presented in different ways for different types of user. Not all reasonable questions will necessarily be amenable to being answered.

9.1.1 Model Form

For those interested in the specification of the models, we should be able to display various components at various levels of detail. This will extend from the top level abstractions applicable to the model, right down to the details of the mathematics involved in the relationships, constraints and distributions in a particular model. Some of this should be shown in mathematical form, but graphical representations should be used wherever possible. For models that fit the Graphical Models framework, the ‘Doodle’ system in WinBugs provides a suitable style of display.

9.1.2 Data used

The model metadata includes links to all the data used in reaching the final calibration of the parameters, so this can be shown, and the user should be able to explore the (separate) metadata for datasets. The links between variables in datasets and those in the model are also available.

9.1.3 Parameters

The final model state includes information about all the posterior distributions, for the (hyper) parameters, for those induced for the parameters of direct interest and for the variables. These can be presented using standard displays of distributions, such as the graphical displays in R.

Such displays show the precision with which parameters have been determined. The reliability and suitability of the model can be explored through the progress of the parameter distributions through the calibration processes.

9.1.4 Domain displays

While generic displays of parameter distributions may be adequate for some statistical users, most practitioners are more used to working with specific forms of display that have been developed as particularly applicable to their domain of application. Here we face the challenge of enhancing such displays to show additional information about (particularly) reliability.

For example, in transport there are specialised displays, such as the network and Origin-Destination diagrams produced by specialised systems such as Visum. It is expected that these can be enhanced to show some aspects of variability and classification, and that they can be used to show appropriate parameter distributions, as well as distributions of actual traffic flows.

9.2 Information Requirements

9.2.1 Basic Areas

We have already identified the three major areas of information about a model that need to be made available to users. These are:

1. The specification of the statistical model that has been fitted.
2. The audit trail of the processes and data used to fit the model.
3. The posterior distributions of the parameters of the model, which contain all the information about the model extracted from the data.

The presentation methods specific to these areas described above will probably be sufficient for the user adept in statistical methods and mathematics. They can use these displays as tools and with them find answers to the questions that they themselves raise about the model.

9.2.2 Domain Users

For a user who is not familiar with statistical methodology (a non-specialist) we need to do more. We will need to provide displays that are simpler (and so do not rely on user understanding of abstract representations), and that are more focussed on the application domain of the user (so we may need different displays for different domains).

The more demanding problem, however, is that we cannot rely on the user being able to formulate appropriate questions, or even recognising that questions need to be asked. So we must address two problems.

1. How to create awareness in the user of the different nature of information obtained from a statistical model, and
2. How to provide a route map for the user through the potentially relevant questions.

For these users it is not enough to provide tools: we must provide solutions, from which they can assess the reliability of conclusions that they may want to draw from the information from the statistical model.

9.2.3 Creating Awareness

Within applications that we control and that provide information from statistical models, we are able to automatically introduce links and prompts to the additional information about provenance and reliability. An example of this in the context of transport networks is the Visum software.

Otherwise we rely on the original authors of the information to create awareness, rather than using software to do it automatically. Where the information from a model is used in an analysis, the analyst should already have made suitable investigations, and so can report these with the analysis and direct the reader to a suitable context for further investigation.

Where information is made available without commentary (and a major example of this is in synthetic datasets), it is necessary to make use of less direct methods, relying on the existence of metadata associated with the information. For example, this is possible for synthetic datasets placed into a Nesstar system, where the DDI metadata allows extensive commentary to be associated with the dataset. It may not be possible in other contexts. In general, we rely on the creator of the information to take every opportunity to direct users to related information about provenance and reliability.

9.2.4 Presentation

Once we have the attention of the user we must guide them to understanding of the nature of statistical models in general, and the reliability of specific information in a particular context.

The approach adopted in the Opus project is to develop a series of web pages that can act as a template for constructing a specific site that would support usage of a group of statistical models within some domain of application.

Some generic requirements for these pages can be identified.

1. They must provide both general guidance and specific information about the model currently of interest. So some pages will be largely static and others will be based on the model metadata.
2. Many users will be interested only in part of the model, probably related to a particular output or component of the underlying system. It should be possible to quickly focus on the relevant parameters and data (the links are in the metadata) without loosing access to appropriate guidance.
3. Particularly in models with large numbers of structured parameters (as, for example, with origin-destination pairs in a transport system) we cannot expect the non-specialist user to explore the whole range. So it is desirable that the presentation system should be able to make some automatic assessment of the reliability of different parameters (or groups of parameters), or the indication of possible anomalies. Further exploration of the Bayesian literature is needed to try to identify suitable measures for this, though we are aware of the inherent danger is such search techniques.

9.3 Evaluation of Model Reliability

9.3.1 Bayesian Model Checking

In all statistical modelling we face the problem of determining whether the chosen statistical model is well-suited to the reality that it represents, and whether it is well-determined by the fitting process that has been used. With classical methods we explore suitability by looking at residuals (comparing observed and fitted data values) and worrying about distributional forms (eg q-q plots), influence, etc. We address quality of fit by looking at measures such as the coefficient of determination (R^2), the residual variance and the significance levels of parameters.

Similar methods can be used in a Bayesian context, though not all have an immediate equivalent. In addition, however, we have the issue that the final posterior distributions are influenced by the initial information in the form and parameters of the prior distributions.

Gelman et al. [GCSR04] discuss model checking in a Bayesian context, and focus on the comparison of data distributions with the equivalent posterior distributions derived from the model. They point out in particular that it is not sufficient to have good correspondence in the mean and variance of a distribution – even with this it is possible to have a poor fit in the tails or inconsistent skewness – so they recommend examining whole distributions to assess model quality.

This is why we place heavy emphasis on the display of distributions as a means to understand model quality and parameter reliability

10. Conclusions:

In this paper we have tried to carry through the following argument:

1. Statistical models are always needed in the analysis of statistical data, and it is better to be explicit about them than to hide them in assumptions.
2. Knowledge and uncertainty about the components or form of a model can be represented by statistical distributions.

3. Bayesian methodology enables us to extract evidence from datasets and use it to update knowledge about the parameters of a model.
4. With multiple datasets related to a modelled system, the Bayesian methodology can be applied multiple times to produce coherent knowledge about the parameters of the model.
5. Statistical models can be difficult to understand, but meta-data about the model specification and fitting processes can be used in presentations that aim to explain the provenance and reliability of results from models to users of these results.

References

[DDI] Data Documentation Initiative. See www.icpsr.umich.edu/DDI for information about the DDI Alliance.

[FWdV03] The Concept of Statistical Meta-Data (2003) by Froeschl, Grossmann, Del Vecchio, a deliverable from the MetaNet project.

[GCSR04] Bayesian Data Analysis. Gelman, Carlin, Stern & Rubin, Chapman Hall, 2004

[GrKe02] The Meta-Data Life Cycle by Ann Green and Jean-Pierre Kent. In Chapter 2 of Deliverable 4: Methodology and Tools (2002), Ed. Jean-Pierre Kent, the MetaNet project.

[LeZo04] Processing, Analyzing, And Archiving Travel Survey Data, by David Levinson and Ewa Zofka. TRB 2005.

[MetaNet] MetaNet: a Network of Excellence for Statistical Meta-data. See www.epros.ed.ac.uk/metanet.

[PPTV01] Modeling Statistical Metadata. Haralambos Papageorgiou, Fragkiskos Pentaris, Eirini Theodorou, Maria Vardaki, Michalis Petrakos: SSDBM 2001

[UML] See www.uml.org for information about UML 2.0. This is a standard developed under the auspices of the Object Management Group (www.omg.org).