Integrating Information about Complex Systems: the Role of Meta-Data in the Acceptability of Results from Models

Andrew Westlake[1] & Marcus Wigan[2]

1 *Corresponding Author*:
Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London
Exhibition Road
London SW7 2AZ
United Kingdom

Phone: +44 20 8374 4723
Fax: +44 87 0055 2953
E-mail: A.Westlake@imperial.ac.uk

2 Napier University Edinburgh, and
Oxford Systematics
GPO Box 126
Heidelberg
Victoria 3084
Australia

Phone: +61 39 459 9671
Fax: +61 39 459 8663
Email: oxsys@optusnet.com.au

Word count: 6541
Figures: 3

11 November 2005

ABSTRACT

The Opus project is developing a methodology for the coherent and consistent integration of information from multiple sources about complex systems, based on Bayesian statistical models. The primary application is in transport. The development of the modelling framework is described elsewhere. Alongside the modelling framework the project has developed a framework for storing a complete audit trail, covering the specification and fitting of the statistical model, in the form of meta-data. This builds on ideas about meta-data for processes that have been developed in other statistical projects.

Practitioners are often sceptical about results derived from models, though in practice all analysis of survey or sample data involves some form of model, even if it is hidden in implicit assumptions. Our approach to this scepticism is to open up the model to scrutiny and to present information about the reliability of results, under the title ‘Provenance and Reliability’. All this comes from the meta-data about the statistical model. While the approach is generic, the presentation needs to be domain-specific and tailored to the level of expertise and understanding of the user of the results.

In this paper we present the motivation behind this approach and describe the meta-data structures and functionality being built to deliver this support to users of results from the Opus methodology.

1 INTRODUCTION

1.1 The Opus project

Opus is a large information management research project, supported by Eurostat as part of the European Commission's Information Society Technologies (IST) Programme. The overall aim of the Opus project is to enable the coherent combination and use of data from disparate, cross-sectoral sources, and so contribute to improved decision making. The research is focused on developing an innovative, generic methodology, incorporating statistical and database systems. Transport planning is a prominent example of a topic that uses multiple sources of data, and is the main test case for Opus.

To meet the needs for comprehensive information on socio-economic systems such as urban and regional transport planning, and in the health services sector, data from diverse sources (e.g. conventional sample surveys, census records, operational data streams and data generated by IST systems themselves) must be combined. There is currently no appropriate developed methodology that enables the combination of complex spatial, temporal and real time data in a statistically coherent fashion. The aim of the project is to develop, apply and evaluate such a methodology. Opus is developing a general statistical framework for combining diverse data sources and has specialised this framework to estimate indicators of mobility such as travel patterns over space and time for different groups of people. The project has pilot and feasibility study applications in London, Zurich, Milan, and on a national level in Belgium.

The benefits of Opus should include:

- Improved estimation of detailed travel demand, using all available information;
- Avoidance of simplified combination of data that can give erroneous estimates;
- Indicators of data quality, to provide guidance for new data collection;
- A framework for managing data from rolling survey programmes;
- Better understanding of the role of variability and uncertainty in results and models;
- Avoidance of confusion from different, apparently conflicting, estimates of the same quantity;
- A generalised methodology for other domains of interest.

The project will present its final results around Easter 2006. Further details can be found on the project web site – www.opus-project.org.

1.2 Modelling with Opus

The Opus methodology is based on statistical modelling, using Bayesian methods to integrate information from multiple sources. In transport (for example) we have specific needs to integrate multiple, partial datasets, but similar problems arise in many domains. Applications always require domain knowledge, and these require specialised models, approaches and assumptions. The project is about the methodology, not the applications, but we treat specific applications as case or feasibility studies in which we explore the problems that arise when the generic methodology is applied. As mentioned, for the project the main test cases are in the Transport domain, with additional feasibility studies in Health.

Opus approaches data integration through the use of statistical models of the domain and problem of interest. The formulation of such models is clearly specific to the application domain and the particular objectives of an analysis. Where multiple, disparate data sources contain information about the domain we use Bayesian methods to combine information about the model extracted from the data sources. Users of results based on statistical models should ask questions about the form and quality of the models used, so we have developed a meta-data-based approach that records the structure of the model and the fitting processes used (as a type of audit trail), together with functionality to present this information to users of the results.

This paper concentrates on the meta-data component of the Opus project, but starts with an outline of the modelling issues, in order to provide context.

1.3 Models in the Opus Methodology

1.3.1 Model Structure

The heart of a model in Opus is a specification in mathematical terms (i.e. largely algebra) of the factors that influence traffic flows (or some other system being studied) and the way in which they interact in their influence. Of course, the particular factors and form of relationships are specific to the problem we are addressing.

All the factors will have statistical distributions associated with them (i.e. they are not necessarily assumed to be fixed), and all the distributions and relationships will have parameters.

All the parameters have prior distributions (representing prior knowledge or uncertainty), which will be more or less informative depending on what experience we can bring to the context and the understanding of the model.

1.3.2 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 of 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 an uncertainty distribution is derived directly from data, we can construct it on any reasonable basis. 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 Bayesian Methodology for working with our models.

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 additional parameters to control the functional form of the model, in addition to those that relate directly to the underlying system.

Very general classes of Bayesian models can be fitted using MCMC (Monte-Carlo Markov Chain) methods. This is the approach used in the Opus project, and details of the fitting methodology are presented in other papers, available from the project web site.

1.3.3 Models and Models

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 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*.

Models are designed to meet a particular need in a particular context. Thus the forms and roles of models can be very varied. Some examples may help to show some of the range.

Conceptual Models are an attempt to form a frame of reference for some domain or collection of constructs or concepts. Where concerned with terminology or names (and so sometimes called *Ontological Models*) they are often similar to classification structures. Other

conceptual models may be concerned with suitable structures for organising ways of thinking about a domain.

The *Relational Database Model* is a formal specification of the structures and behaviour for databases formed from sets of rectangular tables. This provides a conceptual framework for thinking about databases (one that is widely used) but is also sufficiently detailed and precise to be the basis for the implementation of many database software systems.

The *Object Oriented Model* is an alternative (more general) way of thinking about databases and program structures (an alternative paradigm), built using a different set of primitive constructs, assumptions and conventions.

Structural Models concentrate on the objects and attributes that are used to represent information structures. This is necessary for the exchange of information between computer systems, but needs to be accompanied by clear specifications of the intended purpose and use of the various elements. Inconsistent interpretation by independent users or implementers working with such a structure is a continuing concern, unless some enforcement mechanism can be specified and implemented. Structural models can be conceptual, in that they provide a way of thinking about the appropriate structures for some context, or they can be physical, and so present the actual structures needed for some particular system. In contrast, *Process Models* (of which *Data Flow Modelling* is an example) are concerned with the operations that are performed on data objects as they are moved between structures or in response to events. Structural and Process (and related) models for software are often represented using the Unified Modelling Language – UML (see [UML]).

Statistical Models are generally representations of some real system that exhibits variability or unpredictability. They use mathematical relationships to specify the form of dependencies between variables, and statistical distributions to express the variability. Such models can be generic (when they are sometimes described as methods or methodologies), or specific to a particular application (when they will often be instances of the more generic model forms). The term *Graphical Models* is used to refer to a class of models that express dependencies between variables in the form of a graph showing conditional independence. *Bayesian Models* use a formulation of uncertainty about parameter estimates that is based on Bayes' Theorem. For the Opus project we are working with statistical models of specific systems, constructed using the generic Bayesian approach. Some of these models are (in part) instances of Graphical Models.

The term *Transport Model* is used to refer to a class of procedures that are used to estimate information about transport systems that is difficult to observe. For example, 'Route-Flow' models are used to estimate the sharing of traffic flow between possible modes and routes when the demand between origins and destinations is known. Such models are often treated as deterministic, but in reality they are usually statistical models in which variability is ignored. For example, the use of 'least-squares' to estimate the relationship between variables observed with 'error' is only optimal under assumptions of independence, symmetry and constant variance. Many such assumptions can exist within the 'black boxes' that are some transport models. Our preferred approach is to open up such models and to make their mathematical and statistical assumptions explicit.

Models exist at various **levels of abstraction**, and confusion can arise from not recognising the level to which a particular construct contributes, or at which a discussion about the model is taking place. In the Opus methodology we explicitly separate out generalised models (GAPMs – Generalised A-Priori Models) which represent the general knowledge about the nature of relationships and influences within a domain, from the specific and detailed models that are used to explore out understanding or knowledge about a specific issue or system.

2 RESULTS FROM THE OPUS METHODOLOGY

2.1 Why do we use models?

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 statistical 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 so generalise to the whole system.

How, then, do we persuade practitioners that models are valid and useful? One strand of the Opus project addresses this, through the use of meta-data about the statistical models and the model fitting processes. From a philosophical perspective we 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 are concentrating on exposing the qualities of a model so that users can make their own 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.

2.2 The form of Results from Models

The end result from application of the Opus 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. If dealing directly with the mathematics of the model is seen as 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.

2.3 Results from Opus Models

The Opus methodology is Bayesian, so all the knowledge lies in the model specification plus the posterior distributions of the parameters. That is, all information about the underlying real-world system that is contained in observed datasets and is pertinent to the model formulation has already been extracted by the model fitting process into the posterior distributions. In theory it is then sufficient to present just this extracted information (the model formulation together with the posterior distributions) to users. In practice, this will be too complex or impenetrable for most users, so, as with most statistical analyses, other forms of interpretation and presentation will be needed.

Notice that we assume that the mathematical formulation of the model has been determined, and the methodology has been applied to give us the best possible calibration of this model, extracting all possible information from the data sources. Clearly there is a previous process by which the mathematical form of the model is developed and decided upon. This may well use the same methodology as part of an intermediate step (and other methodologies and previous knowledge), but results are always derived from the final version of the model formulation, calibrated in the best possible way.

The central role of the model is valuable because it allow us to generalise, from actual data to all situations covered by the model. All information that we present will be valid information about the model, but will only provide useful insights about the underlying system if the model has a valid (and sensible) structure and is well-determined by the available data. Thus a user of information from the model should reasonably ask about the form of the model, the processes by which it was fitted, and the extent to which conclusions are well-determined.

2.4 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. 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.

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. Later sections of this paper propose 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 – such as synthetic data).

We also need to find ways of presenting this additional information that are accessible and comprehensible for different groups of user. Different types of user will expect answers of different complexity and detail. Some answers can be generic, describing the philosophy behind the Opus methodology and Bayesian modelling, or showing (perhaps in UML diagrams) the outline of the model fitting processes used. 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. The same information may need to be presented in different ways for different types of user.

2.5 The Interpretation of Synthetic Data

Because synthetic data looks like real data, no special facilities are needed to add it to existing data management or analysis systems, or to use it within them. However, synthetic data is different, and to use it effectively (and correctly) the user needs to understand this difference and have access to information about the form and quality of the model.

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.

The issue of sample size illustrates the difference. There can be no information in synthetic data that is not already present in the calibrated model from which it was synthesized, and increasing the size of a synthetic sample just improves the precision of the information about the model, it does not provide any additional information about the real world. With real data, increasing the sample size increases the amount of information available about the real system being observed. However, with simulated data, increasing the size of the simulated sample only provides more information about the model, not about the real system. All available information about the real system has already been extracted into the model.

3 THE ROLE OF META-DATA

3.1 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, including transport. However, the general idea of meta-data has many different applications in different areas and so means different things to 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 [DDI] for the description of survey datasets has achieved wide acceptance, including use by some of the Opus project partners. Several examples of this approach to travel survey data are discussed by Levinson and Zofka [LeZo04].

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.

3.2 Meta-data in Opus

In the Opus project we focus on process meta-data, mostly of extensional form. Our objective is to keep track of the processes that are applied in developing the statistical model from which conclusions are drawn.

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 already presented.

4 THE OPUS META-DATA MODEL

4.1 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 1 shows some of the high level structures that are relevant for this paper. These are developed using the *hyperModel Workbench* [Carl05].

These components are described further in later sections, but first we look at an example. 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.

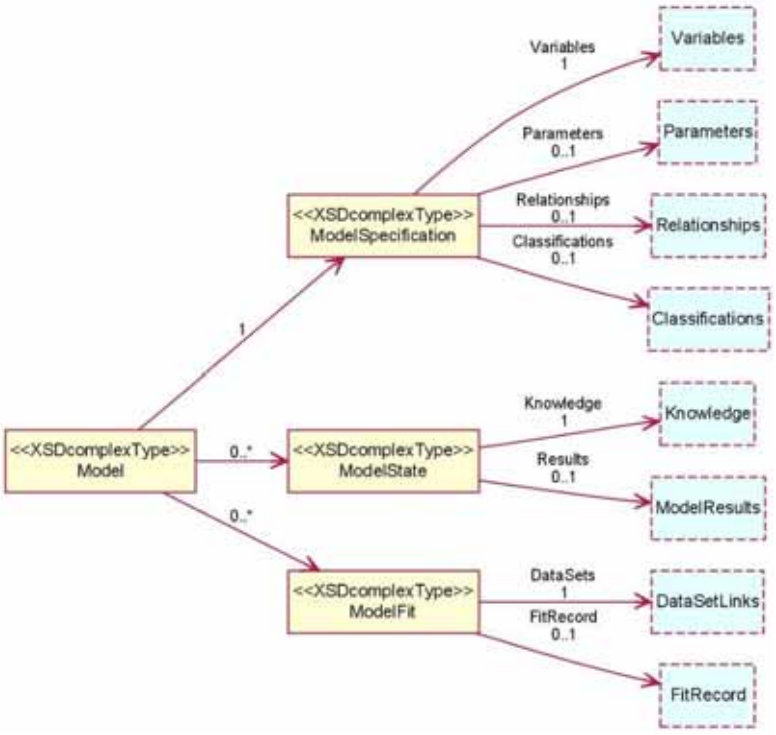

FIGURE 1 Outline of the Opus Meta-data Structure

4.2 Model Instances

Information about actual fitted models is stored in XML documents, with structure controlled by XML schemas generated from the UML structural model. These XML documents are then accessible for use in other software. We are developing a web-based display application to support exploration of models.

FIGURE 2 shows (part of) a web page containing a summary of a particular model. This page is created by using an XML style-sheet to extract and format appropriate parts of the information in the underlying XML document. The layout and the headings (all italicised) are in the style-sheet, and everything else comes from the model information (including the mathematics of the derivations, expressed using MathML).

This example takes two separate matrix estimates of OD flows and fits a common statistical model to them. The model is a main-effects log-linear model (with Poisson variability) for flow between zones, and has separate parameters for within-zone flows.

The display shows most of the textual information available about the model, though it has been truncated here for reasons of space and readability. Following this, FIGURE 3 shows an influence diagram generated from the XML document. This shows how the various relationships (Stochastic, Derived and Constraints) are linked together through the model. This type of diagram is one of the forms of presentation included in the display application that is driven by the XML model specification. The underlying XML document is large, so is not shown here.

Model: **OD Combine**

Combine two separate observations or estimates of numbers of OD trips, fitting a Log-Linear model for the flow rates. Main effects model for the Origin and Destination components, with separate means for intra-zone flows.

Variables:

Name	*Structure*	*Type*
FlowEstimate1	Matrix (*Dimensions*: Origin, *using* OriginZones, Destination, *using* DestZones)	Measure
FlowEstimate2	Matrix (*Dimensions*: Origin, *using* OriginZones, Destination, *using* DestZones)	Measure

Parameters:

Name	*Structure*	
Flow	Matrix (*Dimensions*: OriginZones, DestZones)	The matrix of overall mean OD flows used in the Poisson distributions for observed flows. This is derived from various component influences, in a log-linear model.
FlowLog	Matrix (*Dimensions*: OriginZones, DestZones)	The parameter whose log gives the mean flow. This is defined as a sum of linear components.
FlowAverage	Simple	Factor used for the average (log) number of trips between zones. Should there be separate values for each estimate set?
OriginFlowFactor	Matrix (*Dimensions*: OriginZones)	Factors associated with flow from Origin Zones.
DestFlowFactor	Matrix (*Dimensions*: DestZones)	Factors associated with flow into Destination Zones.
FlowWithin	Matrix (*Dimensions*: OriginZones)	Factors associated with flow within Zones (so Destination is the same as Origin).

Relationships:

Type	*Input*	*Output*	*Form*
Estimate 1 distribution Stochastic	Flow	FlowEstimate1	***Distribution:*** Poisson, *Rate*= Flow Poisson distribution for observations in first estimate set, based on common rates.
Estimate 2 distribution Stochastic	Flow	FlowEstimate2	***Distribution:*** Poisson, *Rate*= Flow Poisson distribution for observations in second estimate set, based on common rates.
Derived	FlowLog	Flow	***Derivation:*** $\exp(FlowLog)$ Poisson rates are derived as exponential of (linear) flow function.
Derived	FlowAverage, OriginFlowFactor, DestFlowFactor, FlowWithin	FlowLog	***Derivation:*** $For: i \in OriginZones, j \in DestZones$ $If: i \neq j \; \eta(FlowAverage + OriginFlowFactor[i] + DestFlowFactor[j])$ $If: i = j \; \eta(FlowWithin[i])$ Linear function for log of flow rates. Inter-zone flow is modelled as an average flow adjusted by origin and destination factors (with no interaction). Intra-zone flows are modelled separately
Constraint	OriginFlowFactor		***Derivation:*** $\sum OriginZoneFactor = 0$ Origin factors sum to zero, so product of rate components is one.
Constraint	DestFlowFactor		***Derivation:*** $\sum DestZoneFactor = 0$ Destination factors sum to zero, so product of rate components is one.

Model States

Prior *ID* = St01. *Type*: Manual. The prior assumptions. These all correspond to almost no knowledge!

Parameter	*Distribution*
FlowAverage	Normal, *Mean*=0, *Precision*=0.001
OriginZoneFactor	Normal, *Mean*=0, *Precision*=0.01
DestZoneFactor	Normal, *Mean*=0, *Precision*=0.01
FlowWithin	Normal, *Mean*=0, *Precision*=0.01

FIGURE 2 Model Summary displayed as a Web Page (incomplete)

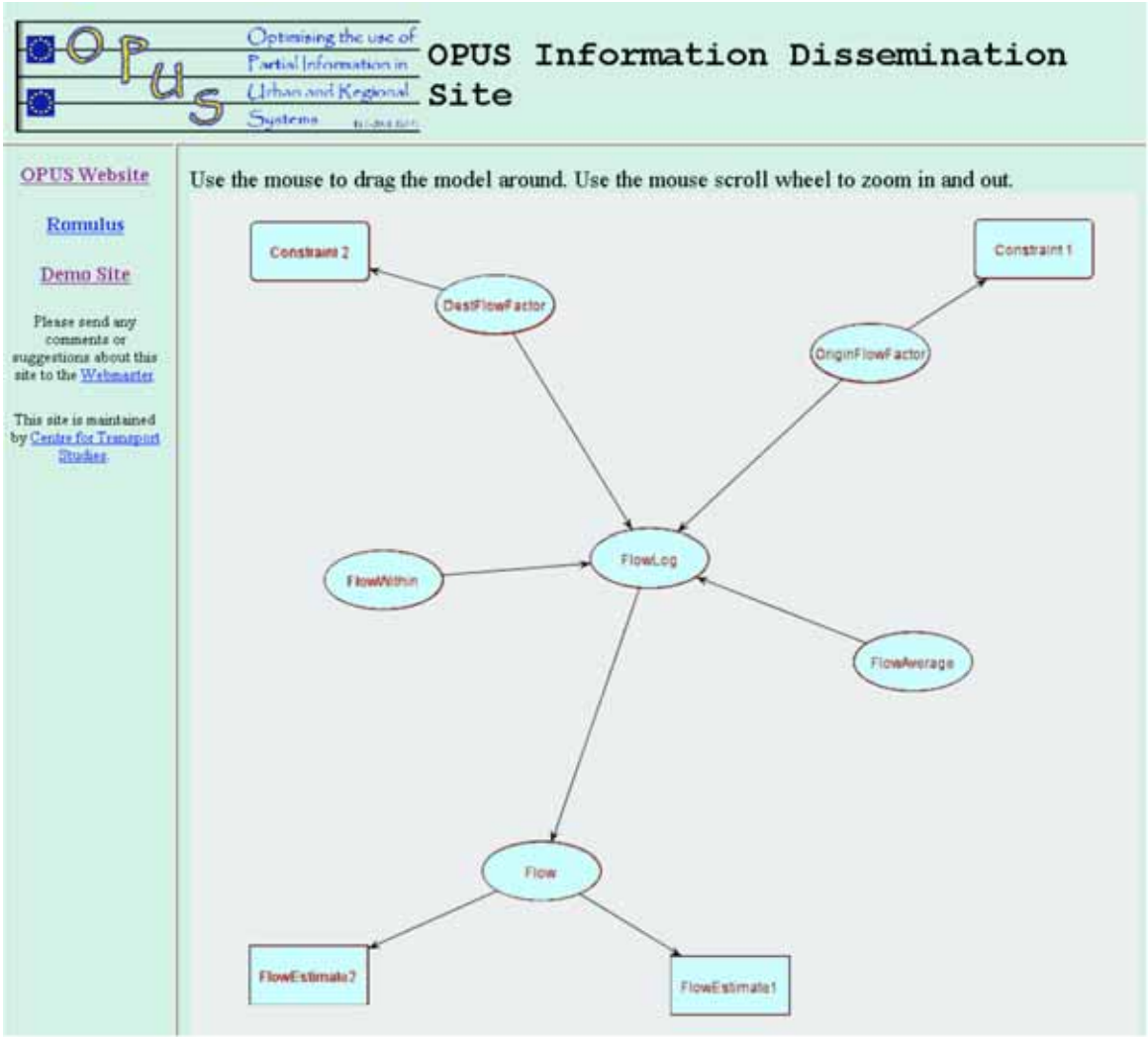

FIGURE 3 Influence Diagram generated from Model Specification

4.3 Components of the Meta-data Model

The **ModelSpecification** is a (single) complex structure that contains all the information about the form of the statistical 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 describes the elements and documents reasons for particular model formulations or parameterisations and for making particular assumptions.

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. When using Bayesian model fitting methods there will usually be an initial state that represents the knowledge brought to the model before any data has been used for calibration. If

we have no prior knowledge then this will be represented by non-informative choices of knowledge distributions.

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 [DDI] for this information.

Where the fitting process is based on MCMC methods, the resulting information about the parameters (the posterior knowledge) takes the form of empirical distributions of simulated values. These are initially stored as multivariate datasets, retaining full information about the distribution and dependencies between parameters, but can also be summarised to (appropriate) specific distributional forms.

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.

4.4 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 that 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, through the meta-data.

4.4.1 Provenance

In this area we focus on the general form of the model specification and the precedents on which it is based (or from which it is derived), plus the datasets used in the fitting steps. It is important to retain the ability for the experienced user to drill down into technical detail, but our initial efforts concentrate on presenting this information at a more general level. Such information will usually need to be presented in a form that is specific to the domain of application and the area to which the model applies. Some more generic presentation may be possible, for example with the generic (GAPM) model diagrams to show influence paths, and with the Graphical Models

that are widely used by statisticians to show conditional independence. Examples of such presentations have been shown in FIGURE 2 and FIGURE 3.

On the model fitting side, we can readily list the steps involved in the fitting processes and the datasets employed in each, together with a brief summary of the contribution of each to the final fit. From here the user can drill down into the meta-data supplied with each dataset.

4.4.2 Reliability

There are two main facets which affect the confidence that a user will have in the results obtained from a model.

The first is the form of the model, where the user needs to be convinced that the model is a reasonable and adequate representation of the aspects of reality to which the results are to be applied. In part this is approached through (detailed) exploration of the model specification, as described in the previous section. It can also be addressed by comparing the distributions of measures in real datasets with the distributions that are predicted by the model for the same measures.

The second facet concerns the extent to which the parameters in the model are well-determined by the fitting processes that have been used. The (posterior) knowledge about the parameters is contained in the final uncertainty distributions, so examination of these reveals the precision of the final knowledge. The knowledge distributions are generally not independent for separate parameters, so multivariate displays (such as contour diagrams) are needed. The statistical package R provides extensive facilities for such displays, so is the target for our development efforts. It also has facilities for the comparison of distributions, for example in different demographic groups or for different origin zones.

Through the mathematical relationships in the model it is possible to derive uncertainty measures for any derivable measure from the model, so this exploration is not limited to the underlying (hyper-) statistical parameters of the model. Because many users will not be familiar with statistical displays we need domain-specific ways of displaying these uncertainty distributions. For example, we are working on ways to represent uncertainty and variability within traffic network flow diagrams.

The sequence of fitting steps provides access to a sequence of uncertainty distributions about any parameter or measure. By examining the way in which the distributions change during the fitting process we can identify where the main changes occur and so which datasets contribute most to the determination of particular measures.

5 CONCLUSIONS

The Opus project is addressing the problem of producing a coherent and consistent view of a complex system through the use of statistical methods to integrate information from multiple sources into a single statistical model. Practitioners are often sceptical about the usefulness of results obtained from models, so we take the initiative to open up the specifications and make information available about the quality of the model.

We have designed a meta-data system which holds all pertinent information about the specification of the model and the processes by which data is used to fit the model. While experienced users may be able to use this information directly, in general we need to be proactive and present information about the provenance and reliability of the model together with actual results obtained from it.

Development and implementation of these ideas is continuing.

ACKNOWLEDGEMENTS

The work reported in this paper (and the whole Opus project) is funded as Project IST-2001-32471, part of the Fifth Framework Information Society Technologies programme of the European Community, managed through Eurostat.

Thanks are due to Opus colleagues Miles Logie and Saikumar Chalisani for the development of the original ideas for this work package, and to Rajesh Krishnan for development work on the presentation application.

REFERENCES

[Carl05] hyperModel Workbench, developed by David A Carlson and others. See www.xmlmodeling.com (last visited 11th November 2005).

[DDI] Data Documentation Initiative. See www.ddialliance.org/ for information about the DDI Alliance (last visited 11th November 2005).

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

[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.

[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).