

OPUS

Optimising the use of Partial information in Urban and regional Systems

Project IST-2001-32471

ITS Programme

Title : **Specification of Pilot Transport
Implementation Model – Inception Report**

Author(s) : Logie, M. (Minnerva)
Lindveld, Ch. (Imperial College London)
Polak, J.W. (Imperial College London)

Deliverable No. : D4.1
Version : 1.3

Date : Initial version: March 2004
Revised: April 2003, May 2004, June 2004

Dissemination Level : LI — Limited to programme participants
Deliverable Nature : RE — Report
Deliverable Type : PD — Programme Deliverable

Project Coordinator : Imperial College London
Contact Person : Prof. John Polak
Address : Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London
London SW7 2AZ
United Kingdom

Telephone : +44-(0)20-7594.6089
Fax : +44-(0)20-7594.6102
e-mail : j.polak@imperial.ac.uk

Consortium : CTS, DEPH, TfL, KATALYSIS, ETHZ, FUNDP,
PTV, SYSTEMATICA, WHO.
MINNERVA, SURVEY AND STATISTICAL
COMPUTING

TABLE OF CONTENTS

TECHNICAL ABSTRACT	1
EXECUTIVE SUMMARY	2
1. INTRODUCTION AND FRAMEWORK	3
1.1 OPUS Project Work Package 4	3
1.1.1 Objectives	3
1.1.2 Description of WP 4 Work	3
1.2 Objectives of Deliverable D4.1	3
1.3 Results presented in Deliverable D4.1	4
1.4 Relation with the OPUS Life-cycle.....	4
1.5 Structure of the Deliverable	4
1.6 About OPUS.....	5
1.6.1 Background and motivation.....	5
1.6.2 Objectives of the OPUS project.....	6
1.6.3 Statistical frame of reference.....	6
1.6.4 Subject areas	7
2. IMPLEMENTATION CONCEPTS.....	8
2.1 Application Objectives	8
2.2 Issues for Implementation	8
2.2.1 Defining an Application	8
2.2.2 Input Data Types	8
2.2.3 Probability Distributions and Measurement Errors.....	9
2.2.4 Input Model(s)	10
2.3 Structure	10
2.3.1 The Role of Structure and Expertise	10
2.3.2 Interactions and Constraints	10
2.3.3 State Variables.....	11
2.3.4 Dimensions	12
2.3.5 Existing Knowledge	12
2.4 Key Ideas from WP2	13
2.4.1 A Transport Generalised A Priori Model (GAPM).....	13
2.4.2 A Health-Transport Generalised A Priori Model (GAPM).....	14
2.4.3 Interpretation of a GAPM in terms of a Graphical Model	16
2.4.4 Summary of GAPM and Graphical Modelling.....	17

3.	IMPLEMENTATION FRAMEWORK	18
3.1	An Overview of the Implementation Framework	18
3.1.1	Definition of Outputs.....	18
3.1.2	Existing Resources	19
3.1.3	GAPM Definition	19
3.1.4	Data and Models: Development of Model-bases	19
3.1.5	Conditional Probabilities	21
3.1.6	Graphical Modelling.....	21
3.1.7	Use of the Complete Model-base	22
3.1.8	Updating	22
4.	OUTLINE EXAMPLE.....	23
4.1	Adapting the General GAPM.....	23
4.1.1	Consideration of Timescales	23
4.1.2	Treating phenomena with shorter time scales than modelled by user equilibrium ...	24
4.1.3	The Simplified GAPM.....	24
5.	CONCLUSIONS	27
6.	REFERENCES	28

TECHNICAL ABSTRACT

This document is Deliverable D4.1 of the Fifth-Framework project OPUS. The OPUS project aims to develop and demonstrate statistically sound methods of combining datasets that each

This deliverable D4.1 is a result of Work Package WP04 of the OPUS project. Work Package WP04 has as title: “Specification of Pilot Transport Implementation Model”. The objectives of this Work Package concerning this deliverable are to

The primary objective of this deliverable is to set the framework for the work being done in WP4. In doing this, it has regard to the following aims:

- Identify the key strands of thinking from WP2 to form inputs to WP4
- Conversely, to provide guidance to WP2 on matters that warrant further attention with respect to concerns of practical implementation and use
- Identify a generic structure and process that can be more fully developed and explained in the results from WP4, namely, deliverable D4.2 Transport Domain Method Specification Report.

The deliverable identifies the general procedure for using MCMC-based graphical models to combine information from data and models to generate enhanced data. This output data is typically in the form of data records describing a synthetic population with characteristics built-up from conditional probability distributions generated from the inputs.

The report defines a framework that is to be further elaborated in the remainder of the work package.

EXECUTIVE SUMMARY

This document is Deliverable D4.1 of the Fifth-Framework project OPUS. The OPUS project aims to develop and demonstrate statistically sound methods of combining datasets that each provide partial or incomplete information on a single complex underlying variable or set of such variables.

The expected practical result of application of the OPUS methodology is a calibrated probabilistic model of the problem domain at hand, with which it is possible to calculate the most likely values of missing, unobserved, or unobservable quantities of the object system under study, with potentially important savings of time and resources.

Objectives of Work Package WP4

The objectives of this work package are:

- To take the generic theoretical framework outlined in WP 2 and to develop the necessary domain-specific and complementary modelling systems and data sources to support a pilot application of the methods in London.
- In particular, to develop systems to enable the estimation of key indicators of urban mobility and the effectiveness of transport policy (e.g., share of trips by different modes of travel, the pattern of origin-destination movements) in London.
- To identify and document the difficulties and problems encountered in the domain-specialisation of the generic theoretical methods, as an input to the refinement of the methods.

Objectives of Deliverable D4.1

The primary objective of this deliverable is to set the framework for the work being done in WP4. In doing this, it has regard to the following aims:

- Identify the key strands of thinking from WP2 to form inputs to WP4
- Conversely, to provide guidance to WP2 on matters that warrant further attention with respect to concerns of practical implementation and use
- Identify a generic structure and process that can be more fully developed and explained in the results from WP4, namely, deliverable D4.2 Transport Domain Method Specification Report.

Results presented in Deliverable D4.1

The following results are presented:

- A structure of a generic flow process that indicates how implementation of the OPUS methodology is to be approached
- Illustrative examples of applying the OPUS methodology to simplified yet representative cases.

1. INTRODUCTION AND FRAMEWORK

1.1 OPUS Project Work Package 4

1.1.1 Objectives

The objectives of Work Package 4 (WP4) are:

- To take the generic theoretical framework proposed in outline in WP2 and to develop the necessary domain-specific and complementary modelling systems and data sources to support a pilot application of the methods in London.
- In particular, to develop systems to enable the estimation of key indicators of urban mobility and the effectiveness of transport policy (e.g., share of trips by different modes of travel, the pattern of origin-destination movements) in London.
- To identify and document the difficulties and problems encountered in the domain-specialisation of the generic theoretical methods, as an input to the refinement of the methods.

1.1.2 Description of WP 4 Work

The initial task of WP4 will be to identify candidate transport data sources and modelling forms that may be used in conjunction with the theory to create a transport-specific methodology. This will be applied in more detail for London and Zurich in the later work packages (WP8 and WP9). The focus of the work will be on assessing the qualities and, if possible, improving the qualities of the models for use with various classes of data. This work includes considering approaches to treatment of modelling errors.

Building on current transport modelling practices, WP4 will specify and demonstrate sets of models that provide information from which the synthesising framework of WP2 can be used to generate synthetic data. In this way, the approach will be to use data to improve models and to use the models to resolve problems of partial or incomplete data. This represents a notable paradigm shift in reducing the distinction between data and models.

An aim of WP4 will be to widen the range of input data that can be used to synthesise transport information. Land use and economic modelling has a role in determining trip frequency. These models have always been associated with transport models, but only recently are being used more regularly. Economic information has relevance also to the synthesis of data on freight movements. Information on traffic and passenger flows from real-time sources is also becoming more available and use of this information is also an objective of WP4.

1.2 Objectives of Deliverable D4.1

The primary objective of this deliverable is to set the framework for the work being done in WP4. In doing this, it will have regard to the following aims:

Identify the key strands of thinking from WP2 to form inputs to WP4

Conversely, to provide guidance to WP2 on matters that warrant further attention with respect to concerns of practical implementation and use

Identify a generic structure and process that can be more fully developed and explained in the results from WP4, namely, deliverable D4.2 Transport Domain Method Specification Report.

WP4 links to a number of other work packages, and it is important that it engages attention and consensus from other work package leaders for the methods that it specifies. This deliverable, and its development, forms an important initial aspect of this process.

1.3 Results presented in Deliverable D4.1

The following results are presented:

- A structure of a generic flow process that indicates how implementation of the OPUS methodology is to be approached
- Illustrative examples of applying the OPUS methodology to simplified yet representative cases.

1.4 Relation with the OPUS Life-cycle

Work Package 4 starts by building on ideas generated from WP2 (Theoretical Framework), but it also contributes to WP2 in the form of an on-going dialogue.

WP4 needs to show how data is to be used and interpreted and therefore needs to work in concert with WP3 (Metadata) and to provide clear direction to WP6 (Database Systems).

The results of WP4 are subject to review and modification by WP5 (Consistency Testing). Where application of processes requires estimation software to be developed, this must be defined by WP4 for use by WP7 (Estimation Software),

The outputs of WP4, as modified by WP5, provide clear guidance to the conduct of the London and Zurich Case Studies (WP8 and WP9 respectively). These case studies will require access to software developed in WP7.

WP4 is focused on the domain of transport information, and the wider implications of its ideas will be studied by WP10 (Feasibility Studies – Transport). WP11 (Feasibility Studies – Health) will show the extent to which the ideas from WP4 are transferable to non-transport domains.

1.5 Structure of the Deliverable

The following Chapter 2 considers the principal issues of concern for translating the theoretical concepts of OPUS into a practical methodology. This includes consideration of how broad abstract ideas can be implemented in varied and complex circumstances.

Chapter 3 defines the key elements of the implementation framework. This chapter provides an important base for development in the main specification deliverable from WP4.

Chapter 4 works through a number of small examples to illustrate the ideas in a more tangible fashion and to establish a *prima facie* case for practical implementation of the OPUS methodologies.

Chapter 5 identifies specific elements of work that need to be undertaken by WP4 and other work packages to meet critical concerns.

The remainder of this Chapter summarises the OPUS project and its objectives.

1.6 About OPUS

1.6.1 Background and motivation

OPUS addresses the situation in which the analyst must combine data from a variety of different data sources to obtain a best estimate, or a fuller understanding, of a system. Such a situation can arise for a number of reasons including:

- No single source contains sufficient information by itself; or
- Multiple sources naturally arise (e.g. through observations at different levels of spatial or temporal aggregation or by means of different survey methods), resulting in a need to reconcile potentially conflicting estimations; or
- The need to update or transfer an existing set of data and parameter estimates when additional information becomes available.
- Problems of combining data from different sources to produce consistent estimates of underlying population parameters arise in many fields of study including environmental monitoring, epidemiology and public health, earth observation, geographic information and navigation systems, transport and logistics, and economic and social statistics. Although the risks of using *ad hoc* combination rules and procedures are well understood, there are nevertheless many examples from practice in which just such approaches are still used. This reflects the fact that, although relatively straightforward methods exist for simple cases, there does not exist a coherent and well developed set of applicable methods capable of dealing with the full range of data combination problems, including factors such as:
 - Data sources that provide both direct and indirect information on the relevant population parameters
 - Data that are presented at different levels of aggregation
 - Data sources with differing levels of statistical precision or user confidence
 - Data that overlap, but that may provide different or conflicting information
 - Gaps in the data observations
 - The issues raised by the aging of sample survey data and the consequent need for updating
 - Accommodating the updating sources
 - The effect of sampling and non-sampling errors (including survey non-response and other sources of missing data)
 - The opportunities presented by new data streams from IST systems

The key scientific objective of the project is to develop a generic statistical framework for the optimal combination of complex spatial and temporal data from survey and non-survey sources. The framework will be sufficiently abstract to be applicable to a wide range of potential domains.

1.6.2 Objectives of the OPUS project

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 overall aim of the proposed project is to develop, apply and evaluate such methodologies, taking as a specific case study the transport planning sector. The specific objectives of the study are:

- To develop a generic statistical framework to enable the optimal combination of complex spatial and temporal data from survey and non-survey sources. This framework will specify how to optimally estimate the underlying population parameters of interest taking into account the structural relationships between the different measured data quantities and the sampling and non-sampling errors associated with the respective data collection processes. It is envisaged that the framework will be broadly Bayesian in nature. The framework will make no specific assumptions regarding the particular structural and sampling/non-sampling errors and will thus be relevant to a wide range of application domains.
- To apply the generic framework within the field of urban and regional transport planning. This will involve the definition of specific structural relationships amongst measured quantities and the characterisation of sampling/non-sampling errors, based on domain knowledge from the field of transport planning.
- To develop the necessary database and estimation software to enable the application of the statistical framework in a number of case study areas.
- To undertake a major pilot application study in London, focusing on the derivation of indicators of the mobility and the performance of transport policy measures.
- In parallel, to investigate the feasibility of applying the framework and methodologies developed both in other transport planning contexts and in other proximate domains, specifically environmental management and social statistics.
- Based on the experience gained in the pilot application and the feasibility studies, to evaluate the performance of the proposed methods and to define the scope and approach for wider applications in relevant domains including environmental management and health care.
- To disseminate the results to the relevant academic and practitioner communities.

1.6.3 Statistical frame of reference

The theoretical approach of OPUS is Bayesian in nature, implying:

An a priori starting point (model) is constructed, including implicit representations of confidence in data sources (through prior distributions) and modelling assumptions;

Additional information is supplied and used to update the model;

The updated model can be used to provide coherent estimators (with the estimates of reliability) for any area that it covers, including combinations of factors for which no data were actually observed. For example it could provide estimates for passengers leaving a

particular railway station in a period when no survey information was collected, but overall passenger loading is known;

As well as parameter estimates, it is possible to use to model to synthesize simulated data sets that demonstrate behaviour of the system, including its variability;

There is scope within the project for the reliance on Bayesian methods to be supplemented with other techniques without altering the general vision. For the present, it is assumed that OPUS will implement its approach using MCMC (Markov Chain Monte Carlo) simulation techniques already widely used in statistical studies, but this is subject to the theoretical phase of work that starts the project.

1.6.4 Subject areas

OPUS provides a generic approach but, in each case, it is necessary to make this approach specific to the particular area of interest (whether the area is geographical or topical in nature). A particular test-bed is transport in London, but studies will be made for transport in Belgium, Switzerland, and Italy, as well as health studies.

2. IMPLEMENTATION CONCEPTS

2.1 Application Objectives

The OPUS methodology, as defined in WP 2, is generic in nature but its implementation for use on any particular application requires that the application and its objectives are defined. The OPUS case studies (WP8 and WP9) and the feasibility studies (WP10 and WP11) are designed to demonstrate the use of the methodology on specific applications with specific objectives. In this work package, we consider example applications which both refer back to examples cited in WP2 and anticipate the interests of the later work packages. In doing this, we consider elements of these example applications, as the intention of the current document (WP4.1) is to define a framework and flow process that is elaborated further in WP4.2 (as previously described in Chapter 1).

The applications that are, somewhat implicitly, used in this document refer to a transport application in which new data is used to update periodic major surveys, and a health-based application in which transport information (in the form of travel patterns for people of different categories) is combined with information on exposure to environmental/pollution hazards. These examples are intended to be illustrative and not to imply constraints on the scope of the OPUS methodology.

2.2 Issues for Implementation

2.2.1 Defining an Application

OPUS has set itself a challenging goal of producing datasets that are rich in information content based on the input of disparate input datasets, as illustrated in a simple manner in Figure 2.1.

Defining an application is therefore a matter of specifying the enhanced output data and the resources, in terms of data, modelling, and expert knowledge on which the enhancement can be based.

2.2.2 Input Data Types

The nature of the processes implied by the ‘OPUS Methodology’ box in Figure 2.1 is clearly dependent on the nature and extent of the data sources provided as input. Complex and varied data sources are bound to imply greater complexity, even though an aim of OPUS is to provide a consistent (Bayesian) approach. The use of more and varied input data nevertheless improves the prospects of deriving rich datasets.

The ability to work with disparate data means that it is possible to gain information from data that is traditionally ignored. In the case of transport modelling, attention is usually restricted to such classes of data such as land-use/planning data, Census (socio-economic) data, travel patterns, and transport infrastructure (network) data. This omits information on person characteristics, as used by market research studies, land and property data, as used by estate agents and property developers, transport condition data as used by travel information

suppliers, and so on, each of which has the potential to enhance the precision of transport modelling.

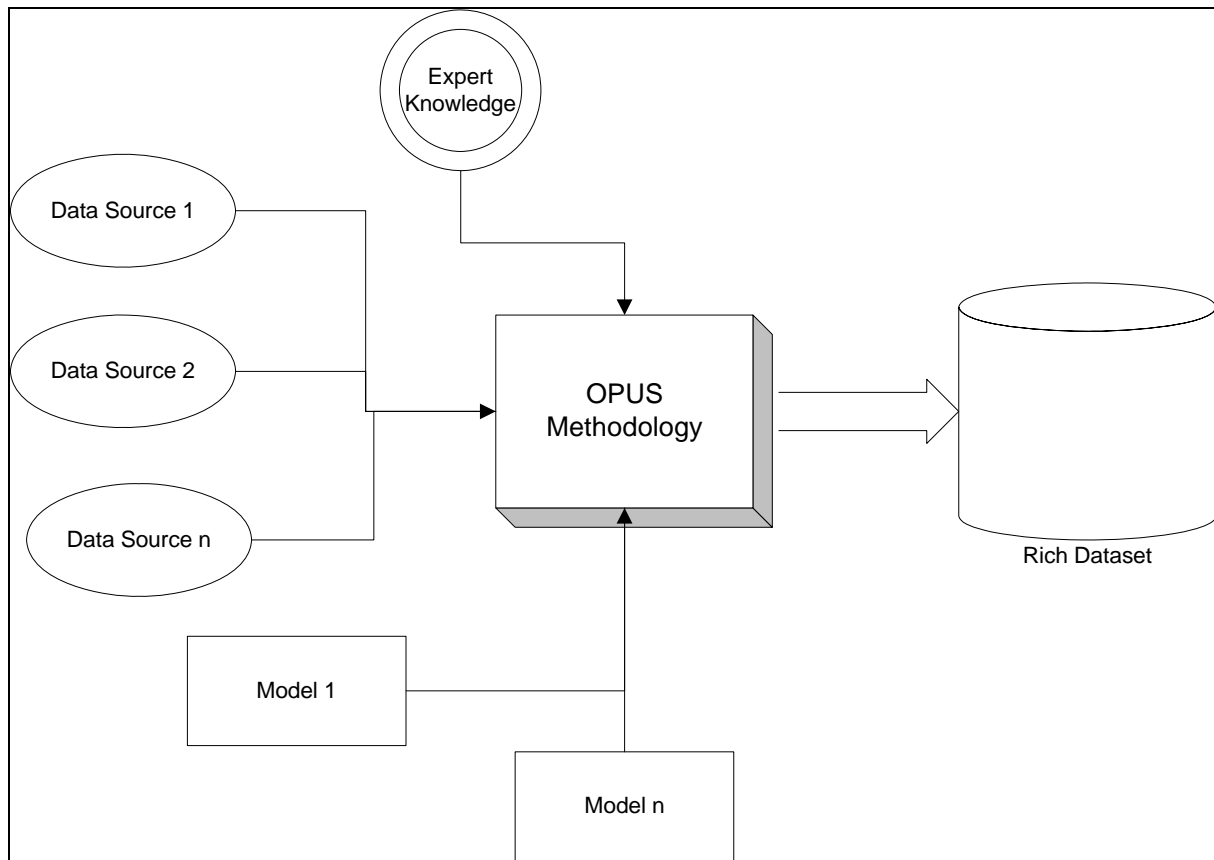


Figure 2-1 Simplified View of OPUS Methodology

By seeking to minimise constraints on input data, except that it should be relevant to the topic domain in question, it soon becomes possible to face the OPUS methodology with great complexity.

It follows that any implementation of the OPUS methodology must be open to inclusion of varied data, but the value of this is limited if the process becomes unwieldy or infeasible to operate.

We consider below ways in which to approach such matters of complexity, while retaining the ability to be open to different and less usual sources of transport-relevant information.

2.2.3 Probability Distributions and Measurement Errors

As is evident from the discussion above, data is considered in terms of probability distribution functions. This is naturally available from disaggregate data sources, but it may be necessary to make assumptions about forms of distributions (e.g. Normal) and of their second moments.

It is also appropriate, and possible within the OPUS framework, to be explicit about measurement and non-response errors associated with survey data. These can be input as probability distributions conditional on the reported data. The specification of measurement and non-response error PDFs may involve assumptions, but this will usually be better than the standard assumption of ignoring these sources of error.

2.2.4 Input Model(s)

From the point of view of the OPUS methodology, models represent a concise way of holding data. Many models output information (data) as mean values, although statistical models often provide more information about the distribution of data.

The modelling inputs to OPUS will typically consist of a set of component models that can be considered together as a single model, and it is therefore a matter of choice as whether to regard this as one or several models.

The ground-work for the approach is provided by the identification of powerful and generally applicable procedures in WP2, but for the moment we concentrate on the role of structure as a means of addressing issues of complexity.

2.3 Structure

2.3.1 The Role of Structure and Expertise

The OPUS methodology needs to offer a structure that can relate, in some way, to all of the data that the user seeks to provide as input. This structure needs to reflect the:

- Dimensions of the data (categorisations)
- Interactions and constraints
- Temporal and spatial variations (also dimensions)
- Existing knowledge, including that expressed in modelling forms.

The OPUS concept is, therefore, to provide a generic structure that may readily be adapted to individual implementations of the methodology.

Defining structure, in the manner further elaborated below, requires expert knowledge. In this respect, OPUS is concerned with exploiting existing knowledge (about data and models) rather than seeking to replicate or replace it.

2.3.2 Interactions and Constraints

The expression of this generic structure is the Generalised A Priori Model (GAPM) that has been introduced in WP2 (Lindveld *et al* (2004)), and which is summarised further below. The GAPM represents a means of identifying the significant interactions and constraints between elements of the system to which the users' data relates.

Some statistical approaches are directed at determining structural information from data, whereas the OPUS approach uses structure as an input. This is primarily based on existing knowledge from established models; the primary requirement is to identify the important structural connections and categories but the methodology will ignore putative input structure that is not sustained by the other input data (that is, the conditional probabilities are found to be negligible).

This implies that the GAPM should be defined with a well-endowed structure. Such a detailed structure implies complexity; this is not a problem at the highest level of abstraction, but for practical implementation it is appropriate to simplify the GAPM to what is deemed relevant and important. Having a concept of a detailed GAPM makes it clearer how different data may

be exploited by the OPUS methodology and how an initial domain of interest might be broadened.

The GAPM provides a modelling framework but, for OPUS, the emphasis is on modelling information flows rather than, necessarily, behavioural characteristics *per se*. This distinction between information processing and behavioural modelling is relevant to determining which interactions and constraints, and hence structure, should be considered important. The resulting view from such consideration provides a means and rationale for combating issues of complexity; primarily by allowing simplification of behavioural issues. In particular, the GAPM can be composed of modelling components that may be considered as ‘black boxes’ by the OPUS methodology, so allowing the user to ignore underlying complexities.

The GAPM represents a starting point for implementing the methodology. A GAPM should be constructed for each topic domain on which the OPUS methodology is to be used. The nature of the GAPM means that it tends towards a holistic view of a topic domain; in this way varied data can be admitted to the process. Implementation of the OPUS method will have a particular focus, according to the input information that is available, and this means that some elements of the GAPM will only be treated in a simplified manner in specific implementations. The simplification may, in practice, mean ignoring elements considered irrelevant but this just means that there are no applicable (observed or modelled) conditional probabilities and they will be ignored anyway by the OPUS methodology. Retaining elements in the GAPM serves to remind that the method is open to further data that might provide conditional probability information.

(The approach of evolving over time and through different applications a far-reaching GAPM for a topic domain provides a convenient way of gaining structural information, and contributes to the efficiency of implementing OPUS methods. However, it is possible to create new, individual GAPMs, especially when the user’s data and perspective is quite different. For example, different GAPMs may be defined for trip-based and activity-based models in the transport domain. However, logically, there should be some connections between even these different perspectives meaning that they could be united in a single GAPM.)

2.3.3 State Variables

The GAPM implies a set of variables used to characterise the state of the system to which the input data applies. The state variables correspond to the inputs and outputs of the modelling components in the GAPM. As with the structure, these state variables may be simplified for particular implementations of the OPUS methodology. Much of the attention of WP4 will be focused on information relating to person trips, which therefore dictates that state variables will link to this concern. An altered focus on, say, freight transport will imply a different, albeit overlapping, set of state variables as will, naturally, different topic domains such as health.

One typical state variable for the transport domain is a ‘trip’, but a matter of possible potential significance for WP4 to consider is whether it is practical to include a ‘tour’ (being a set of linked trips) or indeed the even broader concept of an ‘activity pattern’ in the implementation.

2.3.4 Dimensions

As is familiar from stratified sampling and many modelling processes, better results can be obtained if populations (whether people, trips, etc) can be categorised into reasonably homogeneous groups. Furthermore, we can observe that while different input datasets will support different levels of disaggregation, the output data will be more enriched the greater the level of disaggregation that it supports.

It is therefore important for any implementation of the OPUS method to determine at which level of disaggregation enriched data is to be output. It is possible for the model to be multi-level, with different components relating to different levels.

This is a matter to be investigated as part of WP4, but a target is expected to be provision of data disaggregated to person level giving rise to synthetic samples unifying person and trip data.

Two important aspects affecting the potential for complexity and data enrichment are temporal and spatial dimensions. Spatial dimensions are typically approached in transport modelling via zones, with trips related to paired origin and destination zones meaning that much information is held as two-dimensional matrices. The choice of the number of zones is normally a fundamental early choice for transport modelling, with the amount of computation remaining as a significant consideration.

This will remain the case for OPUS, but it will need to support multiple zoning systems as not all data will conform to a single zoning system.

A characteristic of transport is how it is affected by activities taking place over timescales readily measured in decades (land use changes) to seconds (traffic control). Typical units relate to days or hours, but it will be requirement of any implementation to specify a timescale that is relevant to the input data being used. Nevertheless, OPUS will need to be able to consider data associated with varied timescales and lag effects.

The ability to aggregate spatial and temporal dimensions just described reflects a broader requirement for the OPUS methodology to be able to handle simultaneously data considered at different levels of aggregation.

2.3.5 Existing Knowledge

The Bayesian background to the OPUS methodology serves to underline the value of existing (prior) knowledge, which may be held in diverse forms. One form is as survey datasets, but much information is formally captured as models, so established models are viewed as a source of existing knowledge. This knowledge often concerns structural aspects. For example, travel choice behaviour is frequently represented via hierarchic multinomial logit choice models; the structure and parameters of these models becomes valuable to the OPUS methodology both for use as data and to inform decisions on the structure of a GAPM used for any implementation.

2.4 Key Ideas from WP2

2.4.1 A Transport Generalised A Priori Model (GAPM)

The primary components of a generic GAPM for transport are illustrated in Figure 2.2, which features the key ingredients of:

- Supply and demand
- Interactions
- Constraints

We note that this is not the *only* possible GAPM, and more examples (sometimes for parts of the total GAPM) can be found in Ben-Akiva and Lerman (1984), Fox *et al.* (2003), Cascetta (2001), and e.g. the DIADEM (2003) reports.

It is not a requirement that the GAPM provides an accurate validation but that its behavioural characteristics are plausible. As discussed above, GAPMs include much structural detail at the highest level of abstraction, but practical applications will reduce the detail. For purposes of exposition, Figure 2.2 indicates a relatively simple structure closer to that applying to practical applications, but this should be understood to be derived from a more detailed view. It will be a task of later phases of the OPUS project to define GAPMs more precisely for particular application areas.

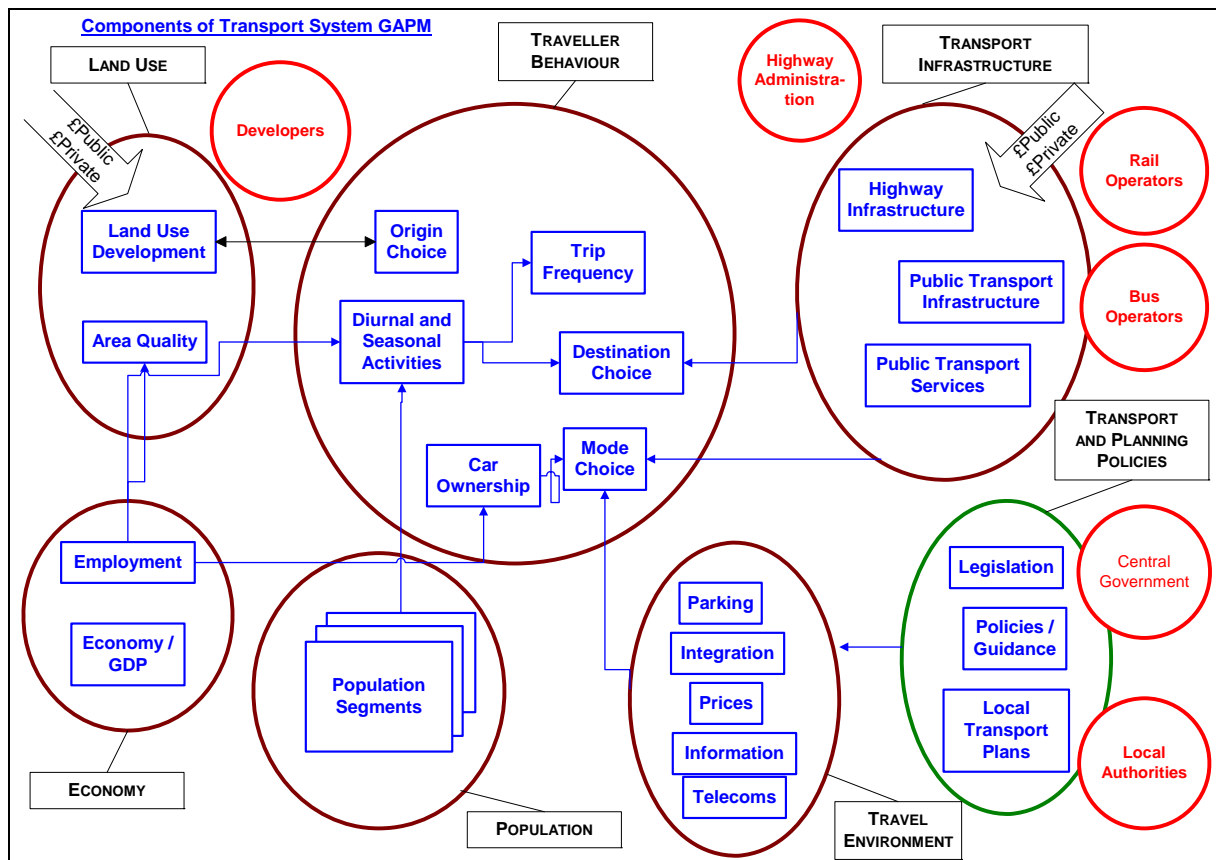


Figure 2-2: A GAPM for transport

It is important for a GAPM to defend what is left as what is kept in. This will be determined by the nature of the inputs; the decisions need to be reflected in the metadata so that it is clear

when introducing new data what changes may be required. This matter is a consideration relevant to WP3.

2.4.2 A Health-Transport Generalised A Priori Model (GAPM)

From the outset, the OPUS project was intended to be applicable across disciplinary boundaries. We will now illustrate how this can be applied to the effect of transport on Health issues through the use of an appropriate GAPM. Once a GAPM has been formulated for this particular problem domain, it can be transformed into a graphical model and be subjected to statistical analysis.

The GAPM will link the transport system with two Health effects: road accidents, the emission of noxious gases, particulate matter, and noise.

We will use diagram in Figure 2-3 to highlight our (transport-centred) view of the apriori relationships underlying the negative effects of traffic on health.

In the top layer we have the autonomous influences of *time* (season, day of week, time of day), and *weather* (temperature, precipitation).

These influence the *transport system* in the middle layer, through seasonal effects and modal split. The transport system (and other sources) determine the *emissions* of type *y* (noxious chemicals, particulate matter, noise) as a function of geographic location *x*, and time *t*. The emissions, the weather, and the number of people in the area determine the overall *Exposure*: $E(n,x,t,s,y)$ where *n* is the number of people, *x*, and *t* are location and time, *s* is the socio-economic population segment, and *y* the type of emission.

At the bottom layer, the exposure rate and a *dose-response model* form one of the factors that lead to disease, although there are numerous extraneous factors. One of the direct effects of the transport system are traffic accidents.

Note that we imply a *stratification* by socio-economic segment, which would allow epidemiologists to study the effects of exposure on, say, children separately from those of grown-ups. This stratification can be obtained from behavioural transportation models, and would be one of the benefits from a closer integrating between the transport model and health-related models.

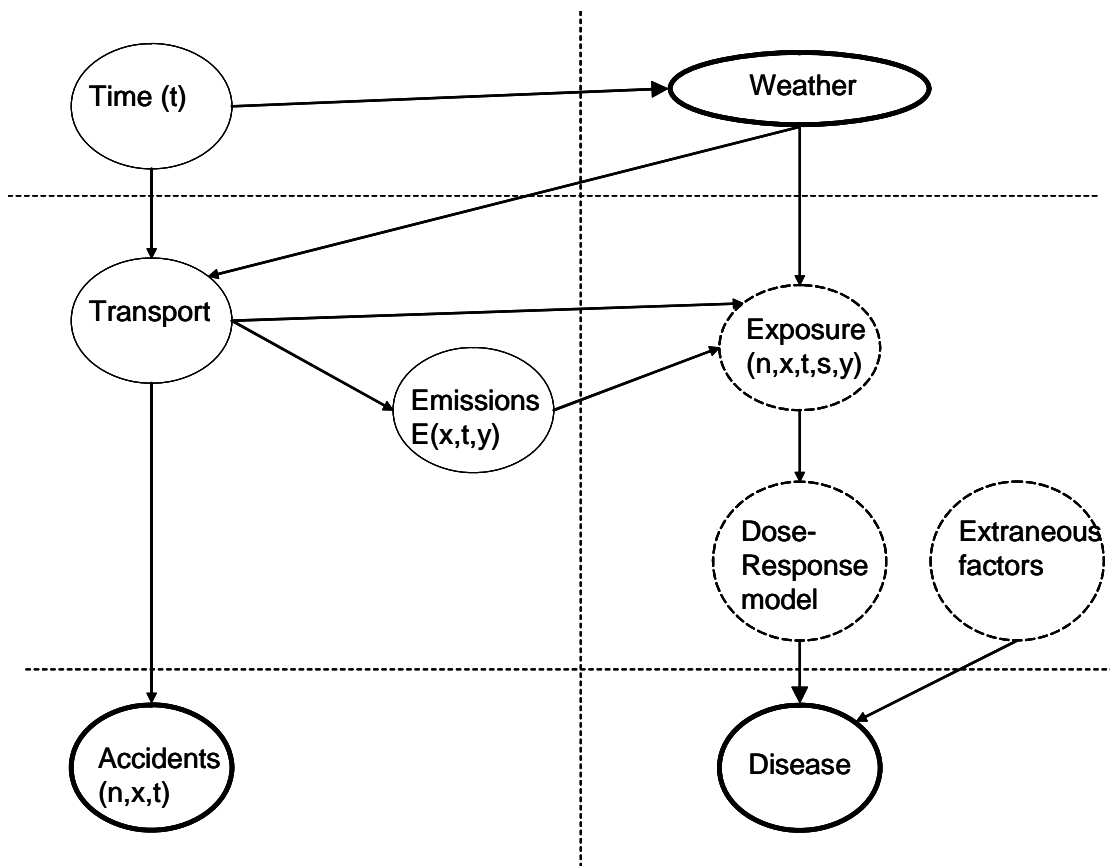


Figure 2-3: General Apriori Model for the negative effects of traffic on health

An expanded version of this figure (especially on the Transport side) is given in Figure 2-4. The expanded figure shows the dependence of mobility on activities, and traffic on mobility. It emphasises that the geographical spread of the population during the day depends on their activity patterns. The influence of time then comes in through its impact on activity patterns. Weather is noted to have an impact on both activities and traffic (modal split). Furthermore the impact of road geometry on emissions (grades) and on traffic accidents is noted. Exposure of the population depends on the number of people on the road (traffic) and stationary (Offices, homes, industry).

The interesting thing is that advanced transport models are able to provide estimates of peoples' socio-economic characteristics s , both in traffic and on location, and hence provide support for dis-aggregation of population exposure rates by socio-economic segment.

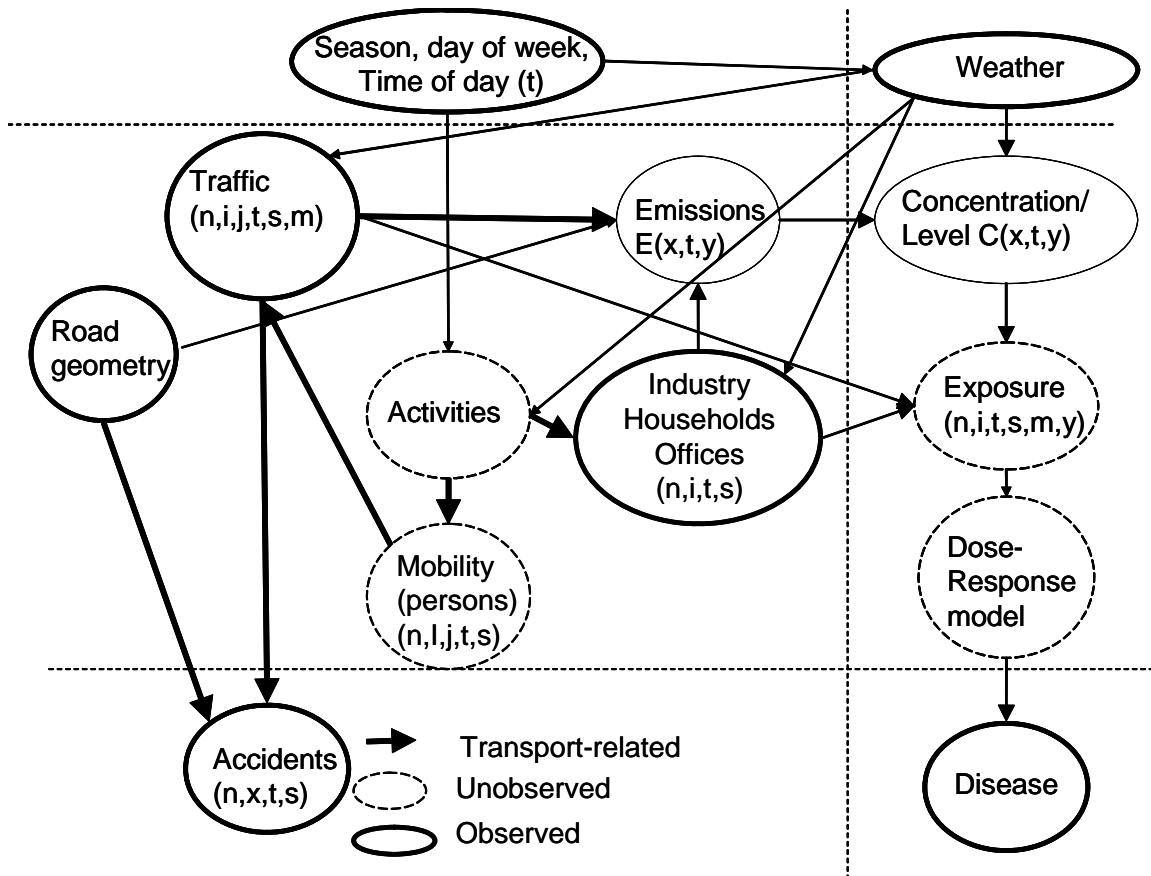


Figure 2-4: General Apriori Model for the negative effects of traffic on health

2.4.3 Interpretation of a GAPM in terms of a Graphical Model

The GAPM proposed in section 2.4.1 covers the levels 2-3 as distinguished in Lindveld *et al* (2004), by specifying:

concepts that are commonly considered relevant for a transportation planning application
their theoretical inter-relationships.

Note for example that the influence of the variable Employment is reflected in the model *only* through the variables to which it is connected (Car Ownership and Diurnal and Seasonal activities), and *no others*. This implies *conditional independence* of all variables in the model on the variable Employment, conditional on the variables ‘Car Ownership’ and ‘Diurnal and Seasonal activities’.

This conditional independence property allows translation of Figure 2-2 into a statistical framework known as a Graphical Model (see Whittaker (1998)) in a relatively straightforward way. In fact, the statement that the conditional independence between variables can be represented as a graph *defines* graphical models.

The natural existence or the ability to make reasonable assumptions of conditional independence is therefore valuable when defining GAPMs as this eases the translation into graphical models. However, it is obviously necessary for the methodology to consider situations when feedback and other effects mean that conditional independence does not

apply between all variables. (In graphical terms this means addressing cases that do not conform to being directed, acyclic graphs.)

In many cases it may be reasonable to consider that such effects may be ignored, either because the input data is adequately contemporaneous and feedback occurs over longer time periods, or because the feedback is a second-order effect. For situation when this is not applicable, say because of the presence of some old input data, it will be necessary for the OPUS methodology to encompass non-acyclic graphs. This is a matter to be pursued further by WP2, but it is likely to involve a two-level process, such as an Expectation-Maximisation (EM) algorithm, in which an assumption of conditional independence applies at one of the levels.

2.4.4 Summary of GAPM and Graphical Modelling

We can summarise the considerations in the discussion above as follows:

- In order to re-use and leverage existing tools and solutions, similarities between the problem instances that were already solved and the ones that we need to solve should be found
- The OPUS methodology will limit itself to subject areas for which an adequate body of theory is available, preferably in the form of Generic Apriori models (GAPMs).
- The use of GAPMs largely eliminates the need to infer the model structure from the data, and leads to models that are much likely to be reasonable, correct, and capable of being estimated efficiently than models that are not based on GAPMs.
- In order to codify the GAPMs in a uniform way and thereby to allow comparison, we propose to express them in terms of Graphical Models
- Graphical Models seem well-suited to the tasks of:
 - Capture the essential structure of GAPMs
 - Manage the complexity of GAPMs
 - Capture model-uncertainty in the GAPMs

An element of the general approach should therefore be to translate the GAPM into a Graphical Model, and to rephrase the question for simultaneous use of different datasets in terms of the Graphical Model.

3. IMPLEMENTATION FRAMEWORK

3.1 An Overview of the Implementation Framework

In this Chapter we bring together the concepts discussed in this document to define a framework for implementing the OPUS methodology. This framework is essentially a process for guiding the user through a series of logical steps that transform and unify different sets of data into an enriched set.

The specification of this process is the primary concern of the final deliverable from WP4, but here we identify the main elements of the process.

The implementation framework is shown schematically in Figure 3.1, which we interpret further as follows. For illustration, the discussion uses examples from the topic domain of transport in London, being relevant to WP8, the London Case Study.

3.1.1 Definition of Outputs

Outputs for transport-based applications of OPUS will generally be distributions of information on people, their characteristics, and their travel. This may logically be represented as multi-dimensional cubes, but a typical form of output will be as data records ('trip records' in transport modelling terminology). The output will therefore be disaggregate data sets corresponding to 'synthetic populations', with the choice of input data determining which population is applicable.

In transport work the broadest population is normally provided by traveller surveys, so the resulting synthetic population would be of people who travel. Traveller surveys provide information about trips, but little about the type of people making them or the relation with other trips (as part of a 'tour'). Providing this additional detail to the trip records represents one type of OPUS output.

Because information is available about people in the synthetic population, it is possible to add detail about the people, say cultural and life-cycle preferences, which can aid transport studies or information on health characteristics, which can aid health studies interested in location-based activities such as travel.

Additional detail may relate to a specific dimension, such as time. This can be relevant when the identified input data sources include automated measurement sources (as are becoming increasingly available via IT-based systems) that produce detailed and comprehensive time-based information.

In other applications, the scope of the population may be expanded, rather than adding detail to a particular population. For example, Census data might be used to provide information on non-residents (visitors) to household survey populations.

It is not necessary for the population details or extent to change; OPUS can be applied in cases where the input data includes newer observations of the same population to provide an updated view.

Although we focus here on transport information, it is clear that that this is not a requirement of the OPUS methodology.

3.1.2 Existing Resources

Once a required output has been identified, the starting point for applying the OPUS methodology is to identify the existing resources for the domain of interest in terms of data sources and established models. Data sources will typically comprise survey data (Figure 3.1 identifies the London ‘LATS’ survey data as an example input, which itself comprises varied surveys including household, roadside, and passenger travel surveys), but non-survey data sources, such as administrative information, or data that is a by-product from other sources (e.g. Congestion Charging, as an example in London) are also of value.

We refer to these existing data and modelling resources as the ‘primary inputs’.

3.1.3 GAPM Definition

A starting point for organising the primary inputs is to construct a GAPM that identifies interacting processes that connect the various sources of data. The process of constructing a GAPM is based on adapting an existing GAPM for the domain in question – this is on the basis that there are generic processes and interactions relevant to a domain, but that individual implementations will need to reflect specific issues and features. The adaptation is essentially adding any components required by the data and not previously considered in the existing GAPM. For simplicity, rather than necessity, redundant elements may be removed.

3.1.4 Data and Models: Development of Model-bases

Once a set of primary inputs has been identified as available and relevant, the data and modelling information is grouped into what are termed ‘model-bases’, that is, databases that include both data and modelling information.

These model-bases are represented in Figure 3.1 as cubes to indicate the multi-dimensional nature of the data that they hold.

Because the OPUS methodology is open to different data and models, the contents of the model-bases may be quite varied. They will typically involve components of:

- Data, suitably but not necessarily, held as relational tables
- Model parameters, typically held as scalar and vector sets
- Model structure, typically represented as a tree-structure or similar.

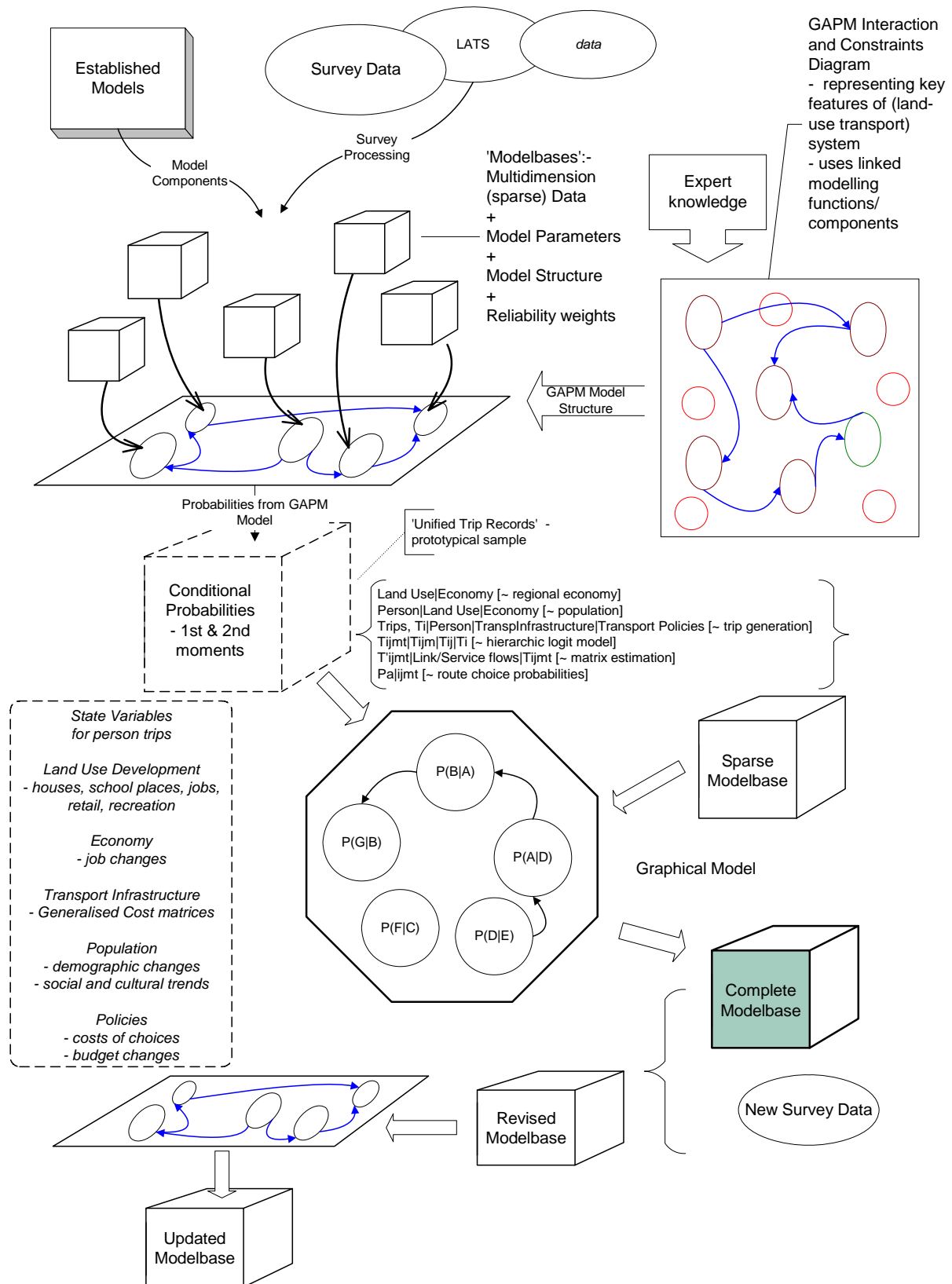


Figure 3-1 Schematic Implementation Framework

The open nature of the model-bases requires that the model-bases are 'self-explanatory' to the processes that will use them. This is to be accomplished through the provision of metadata associated with each model-base whose design is the subject of WP3.

One aspect of this metadata is to provide information on the relative reliability of different sets of information. This is done in the form of weights which are set relative to a benchmark set of data (determined by the user). These weights are classified as metadata as they describe the input data, but they are used as data in the methodology.

3.1.5 Conditional Probabilities

The linkages defined in the GAPM indicate where processes have an influence on each other, which may be understood in terms of input-output diagrams, a simple element of which is shown in Figure 3.2.

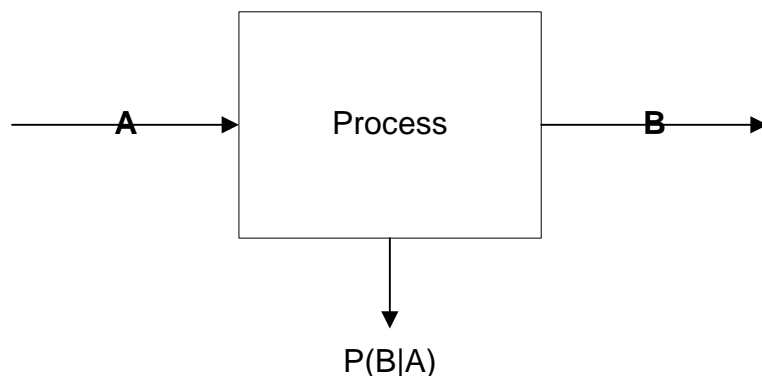


Figure 3-2 Simple Input-Output

In this view, the processes can be seen as a means of generating conditional probabilities. Most models in the transport domain provide information on the first moment of the probability distribution, although by examination of the underlying assumptions used to model the process, it is often possible to infer a second moment. Some elements of the GAPM will relate to stochastic and simulation modelling which will offer information on second moments more directly.

A standard trip-based transport GAPM will give rise to a number of sets of conditional probabilities related to different components of the GAPM, as illustrated in Table 3.1.

Table 3.1 indicates conditional probabilities in terms of example state variables (e.g. Land Use, Person Type, etc). It uses standard notation for trip state variables, thus T_{ijmt} represents a trip between zones i and j , using mode m at time period t .

3.1.6 Graphical Modelling

Using the concepts introduced in WP2, Lindveld *et al* (2004), the GAPM represents a Mathematical Model that can be translated into a Statistical Model, which is provided in the OPUS Methodology by a Graphical model.

The result of the GAPM is therefore to enable the construction of a Graphical model with an available set of conditional probabilities, generally with associated first and second moments.

The Graphical model provides the mechanism to generate more complete sets of conditional probabilities from the partial/sparse data that is provided by the GAPM. These fuller sets of

conditional probabilities, when coupled with the (relatively) sparse information in the model-bases allows fuller sets of information to be generated that become the OPUS output in the form of a complete model-base.

Table 3-1 Identification of Conditional Probabilities from GAPM Components

Conditional Probability	GAPM Component/Model
[Land Use Economy]	Regional economy
[Person Type Land Use, Economy]	Population demographics
[Trips, T_i Person, Transport Infrastructure, Transport Policies]	Trip generation
[T_{ijmt} T_{ijm} , T_{ij} , T_i]	Travel behaviour (hierarchical logit model)
[T'_{ijmt} Link/Service flows, T_{ijmt}]	Matrix Estimation
[$P_{a ijmt}$]	Route choice probabilities (Assignment)

3.1.7 Use of the Complete Model-base

The quality of information in the Complete, output Model-base will vary according to the precision and extent of the primary input information. It is therefore important that output model-bases are also supplied with metadata that informs the user about the reliability of the information. This information particularly needs to guide the user as to the level of aggregation across dimensions for which it is safe to use the outputs. This aspect is therefore also an important element of WP3.

3.1.8 Updating

The methodology for implementation is suited not only to generating a set of richer information, but also of allowing this data to be kept up to date with the supply of new primary inputs. New data can extend and supplant information in the previous model-bases, providing the basis for re-execution of the process to generate a new, refreshed output Model-base. This process needs to include updated metadata to reflect changes in data reliability arising from the passage of time and other causes.

4. OUTLINE EXAMPLE

This chapter aims to further explain the ideas presented previously through their application to a simple example. The priori GAPM presented in Figure 2-2 contains a lot of complexity that will not be used in each and every application. One of the issues is how to adapt a generic GAPM to the situation at hand.

4.1 Adapting the General GAPM

The priori GAPM presented in Figure 2-2 contains some complexity that will not be used in each and every application. One of the issues is how to adapt a generic GAPM to the situation at hand.

The generic GAPM can be adapted to the situation at hand by judiciously omitting specific interactions, thereby reducing its complexity and making it easier to use.

One of the ways in which the generic GAPM may be reduced is by noting that many of the elements presented in it operate on different time-scales. As noted e.g. in Ben-Akiva and Lerman (1984), a *choice hierarchy* can be distinguished in transportation systems according to the time scale and the transaction costs of the phenomena considered.

For example traffic flows depend on route-choice decisions, which can typically be revised in hours or days. The choice of a residential location on the other hand usually remains fixed for years. Therefore a model with a focus on within-day traffic dynamics may be able to treat all decisions of route and mode, let alone residential choice and destination choice as *exogenous*.

4.1.1 Consideration of Timescales

The approach of treating phenomena with longer timescales than modelled as exogenous is illustrated in Figure 4-1, where the time-scale of events is set out on the horizontal axis, and ranges from decades in the case of land-use and population changes to days or hours as regards the characteristics of the traffic flow. Assumed is a modelling horizon of medium to short-term events, so that the long-term characteristics presented by housing, jobs, and constitution of the population etc. can be assumed given.

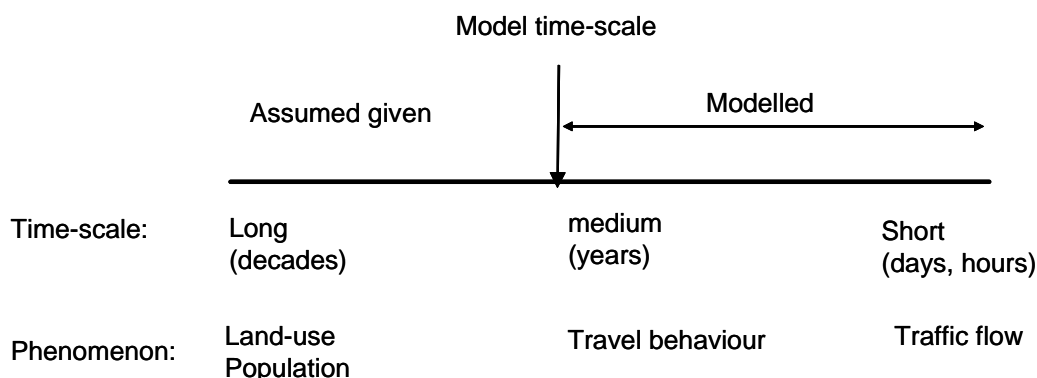


Figure 4-1: Long-term input considered fixed

4.1.2 Treating phenomena with shorter time scales than modelled by user equilibrium

On the other hand a model which focuses on the interaction between land-use and mobility may be able to assume that all shorter-lived decisions are made in such a way that the hypothesis of *user-equilibrium* is valid. This can be presented in as in Figure4-2:

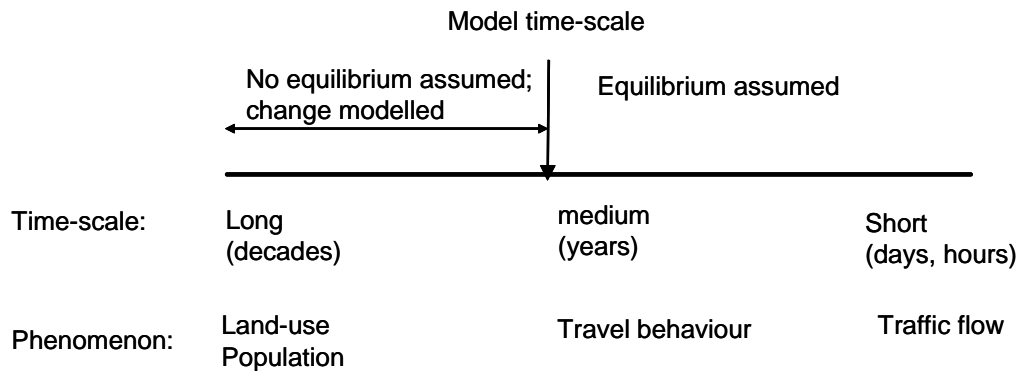


Figure 4-2: Decision Hierarchy

4.1.3 The Simplified GAPM

With these considerations in mind, we can simplify the GAPM as follows:

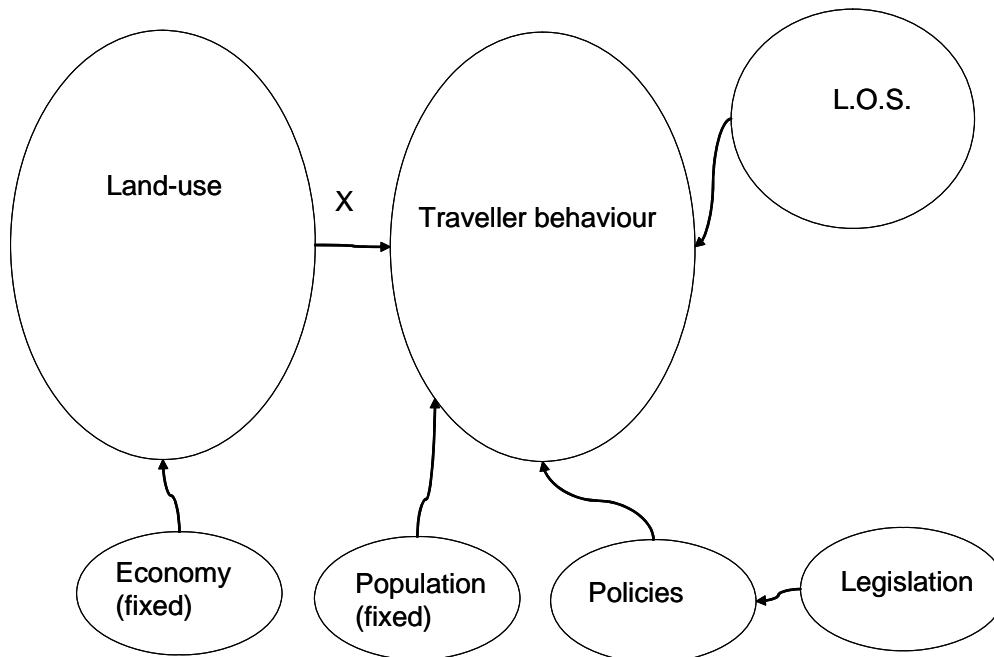


Figure 4-3: Reduced GAPM

Where the vector X consists of a set of land use and socio-economic attributes (such as employment and residential density and income) that travel behaviour, represented by the vector Y .

A graphical model of updating this situation could be:

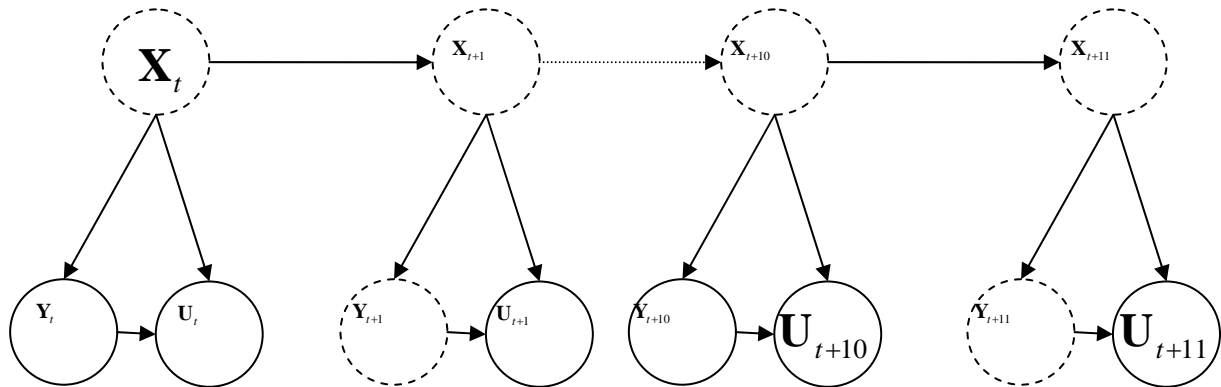


Figure 4-4: Graphical model of updating land-use through 2 surveys

Denoting the (latent) vector describing land-use by \mathbf{X}_t , we note that it develops over time over a period T , giving a time-series of vectors: $\{\mathbf{X}_t\}_{t \in T}$. We will assume that every k years (we assume $k=10$ years for the moment) a large-scale (disaggregate) survey is conducted, giving an observed data vector Y_t . We also assume that each year up-to-date but aggregate observations are available: U_t .

The structure shown in Figure 4-4 is reminiscent of that of a Kalman filter, especially if we consider each of the surveys separately.

However, the complication that the aggregate survey U_t does not add much information if the disaggregate survey Y_t is available, as is shown in the following part of the graph:

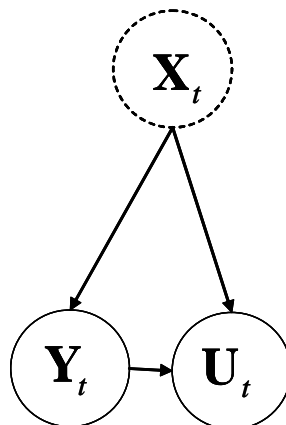


Figure 4-5: Latent land-use, disaggregate and aggregate surveys

Given that U_t and Y_t are not independent we have to take account of their correlation. In a Kalman filter we would simply extend the observation vector to $\begin{pmatrix} Y_t \\ U_t \end{pmatrix}$, use this to update with

a frequency of once per 10 years, and otherwise update using U_t alone. The variance-covariance matrix corresponding to $\begin{pmatrix} Y_t \\ U_t \end{pmatrix}$ would be used to weigh the relative contributions of U_t and Y_t .

The next (and more serious) problem is that distributional assumptions (normality) that underlie the Kalman filter do not seem appropriate. If we wish freedom from the assumption of normality, the Kalman filter loses its optimality, and the exact expressions based on graphical modelling should be used instead.

Noting that the graph in Figure 4-5 directed and acyclic, we can use MCMC methods to calculate the probability distribution $P(X_t | Y_t, U_t)$.

We also note that the disaggregate survey data Y_t does not immediately become worthless one year after it is collected. Rather it will lose weight relative to the more up-to-date data in the aggregate survey U_t . The relative weights of U_t and Y_t can be thought to follow the sawtooth profile as shown in Figure 4-6.

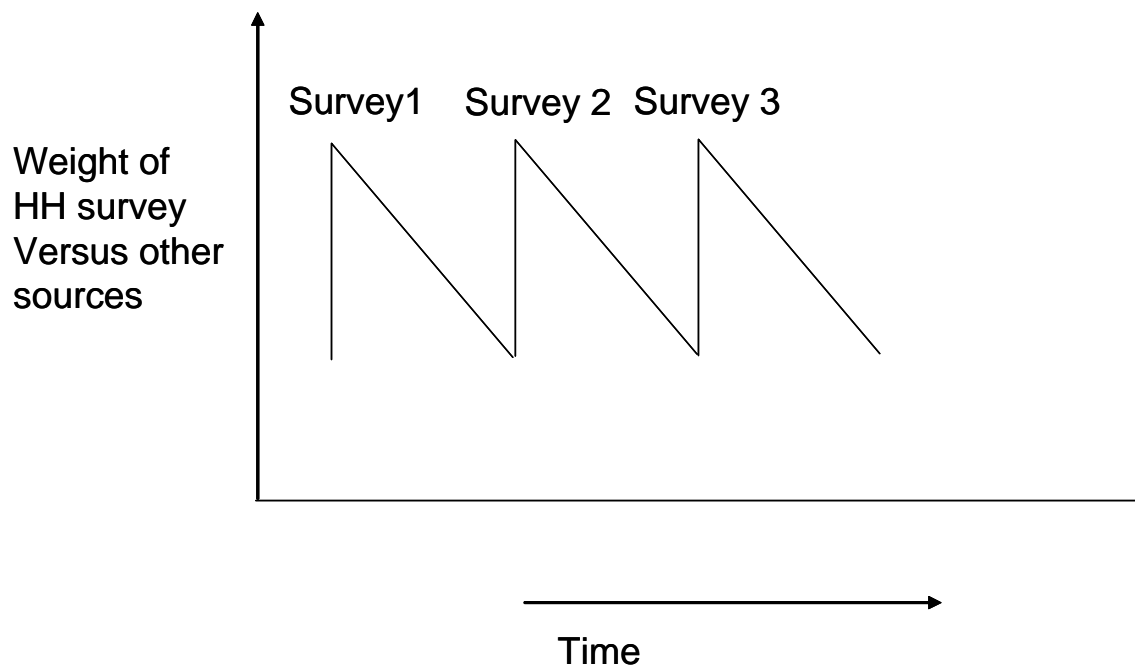


Figure 4-6: Sawtooth profile of relative weights of aggregate and disaggregate surveys

5. CONCLUSIONS

In the Deliverable we have begun the process of translating the theoretical concepts developed in WP2 into a transport domain application. We have considered how existing domain knowledge regarding both underlying structure and measurement processes can be interpreted within the formalism of Bayesian graphical model, explored how relevant problems might be represented and discussed some of the practical considerations associated with implementation.

These themes will be taken up and developed in more detail during the course of the further work in Workpackage 4.

6. REFERENCES

Ben-Akiva, M., Lerman, S.R. (1984) *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT press, Cambridge.

Cascetta, E. (2001) *Transportation Systems Engineering: Theory and Methods*. Kluwer Academic Publishers.

DIADEM (2003) UK Department for Transport Seminar Papers on Variable Demand Modelling and DIADEM, 17th July 2003, Great Minster House, London. URL: <http://www.dft.gov.uk/> and use search engine on keyword diadem.

Fox, J., Daly, A. and Gunn, H. (2003) Review of RAND Europe's Transport Demand Model Systems. Report Prepared for TRL Limited. RAND. <http://www.rand.org/publications/MR/MR1694/MR1694.pdf>

Lindveld, Ch, Collop, M. Logie, M, Polak, J.W. and Westlake, A. (2004) *Identification of Methodology and Tools*, OPUS Deliverable D2.1. <http://www.opus-project.org/>

Logie, M. (2003) *The OPUS vision*. Internal note. <http://www.opus-project.org/>

Ortuzar, J.de Dios and Willumsen, L. (2001) *Modelling Transport*. Third edition. John Wiley.