

# OPUS

## *Optimising the use of Partial information in Urban and regional Systems*

**Project IST-2001-32471**

**ITS Programme**

**TRANSPORT**

**Title :** **Identification of Methodology and tools**

**Author(s) :** Lindveld, Ch. (CTS)  
Collop, M. (TfL)  
Logie, M. (Minnerva)  
Polak, J. (CTS)  
Westlake, A. (Survey and Statistical Computing)

**Deliverable No. :** D2.2  
**Version :** 0.6

**Contract Date :** April 2003  
**Submission Date :** April 15, 2004

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

**Project Coordinator :** Imperial College London  
**Contact Person :** John Polak  
**Address :** Imperial College  
South Kensington campus  
London SW7 2AZ  
United Kingdom

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

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



---

## TABLE OF CONTENTS

---

<b>Technical Abstract</b>	<b>6</b>
<b>Executive Summary</b>	<b>7</b>
<b>1. Introduction and Framework</b>	<b>8</b>
<b>1.1 The OPUS project</b>	<b>8</b>
<b>1.2 Objectives of the OPUS project</b>	<b>9</b>
<b>1.3 Background and motivation</b>	<b>10</b>
<b>1.4 Statistical frame of reference</b>	<b>10</b>
<b>1.5 Subject areas</b>	<b>11</b>
<b>1.6 Relation with the OPUS life-cycle</b>	<b>11</b>
<b>1.7 Objective of the Deliverable</b>	<b>11</b>
<b>1.8 Structure of the deliverable</b>	<b>12</b>
<b>2. Problem instances</b>	<b>13</b>
<b>2.1 Transport</b>	<b>13</b>
2.1.1 Transport modelling	13
2.1.2 The use of prototypical samples in transportation demand models	14
2.1.3 Estimation of travel demand (mobility)	14
2.1.3.1 General approach	15
2.1.3.2 Natural Variation	15
2.1.3.3 Measurement Bias	16
2.1.3.4 Indirect Measurements	17
2.1.3.5 Non-response Effects	18
2.1.3.6 Formulation of the Likelihood Function	18
2.1.3.7 Scope of application	19
2.1.4 Combined SP-RP estimation	21
2.1.5 Positioning	22
2.1.5.1 Background	22
2.1.5.2 The transition equations	22
2.1.5.3 The observation equations	23
2.1.5.4 Kalman Filter Design	24
<b>2.2 Health</b>	<b>26</b>
2.2.1 Example	28
2.2.2 Simulation study	29

<b>3. General approach</b>	<b>30</b>
<b>3.1 Conceptual hierarchy</b>	<b>30</b>
3.1.1 An example: O-D matrix estimation	31
<b>3.2 Finding commonalities and discrepancies</b>	<b>32</b>
<b>3.3 Generic methods and specific instances</b>	<b>32</b>
3.3.1 Constructing a starting point	32
3.3.2 A transport General A Priori Model (GAPM)	32
3.3.3 Interpretation of a GAPM in terms of a graphical model	33
<b>3.4 General approach proposed</b>	<b>34</b>
<b>4. Formal methods and software</b>	<b>35</b>
<b>4.1 Literature scan</b>	<b>35</b>
<b>4.2 Representing the GAPMs of the problem domain</b>	<b>36</b>
4.2.1 Structural equations models	36
4.2.2 Graphical models	36
4.2.2.1 Graphical models and conditional independence	38
4.2.2.2 Separating measurement relationships from structural relationships	38
4.2.2.3 A succinct characterisation of graphical models	39
4.2.2.4 Operationalising graphical models	39
4.2.2.5 Pro and contra of graphical models	39
<b>4.3 Identification of probability distributions of observable quantities</b>	<b>40</b>
4.3.1 Partial identification of probability distributions	40
4.3.1.1 Relevance of the work on partial identification	41
4.3.2 Analysis of incomplete data under the MAR hypothesis	41
<b>4.4 Filtering</b>	<b>42</b>
4.4.1 The Kalman filter	42
<b>4.5 Software identified for use</b>	<b>44</b>
4.5.1 General technical computations and prototyping	44
4.5.2 Statistical computations	44
4.5.3 Graphical modelling	44
<b>5. Conclusions</b>	<b>45</b>
<b>5.1 Commonalities noted between successful applications</b>	<b>45</b>
<b>5.2 Partial solutions identified</b>	<b>45</b>
<b>5.3 Contours of a unified framework</b>	<b>46</b>
<b>6. References</b>	<b>47</b>



## TECHNICAL ABSTRACT

---

This deliverable D2.2 is a result of Work Package WP02 of the OPUS project. Work Package WP02 has as title: “Theoretical Framework”. The objectives of this Work Package concerning this deliverable is identify and outline the formal methodology and the tools to be used in combination of datasets in the OPUS project. Although at this stage of the project the theoretical framework obviously cannot be complete, its outlines and guiding principles seem to have solidified.

A number of successful examples of combining datasets have been found in the literature, along with a few unused opportunities for combining datasets in the problem domains directly addressed by the demonstration projects within OPUS.

At first sight the successful examples are different from the cases where the OPUS project aims to apply the merging of datasets. The issue is how to port the ideas and techniques that underlie these successful applications in their respective fields to the domains that OPUS is concentrating on in a sufficiently generic way. It was then argued that the specifics of the problem domain obscure this process, and that similarities could be expected only at the level of mathematical and statistical descriptions of the problems concerned, leading directly to a relatively abstract and mathematical approach.

The underlying principle of merging datasets in a statistical way seems to be the construction of a joint probability distribution of all relevant variables in the model of the problem domain. With this distribution in hand, and given a suitable parameterisation of the models, a likelihood function can be constructed that contains all model parameters and uses all datasets. Determination of the parameters from such a likelihood function automatically combines the datasets in a statistically sound fashion.

However, two significant practical problems emerge: how can one construct such a complicated likelihood function in a manageable and practical way, and how can one use it (e.g. to determine parameter values) if it is not analytically tractable (which is clearly not guaranteed).

The first question is answered partly by an appeal to accepted *theory of the problem domain* as represented in a Generic APriori Model (GAPM; *assumed to be available*) and partly by the translation of such a GAPM into a graphical model based on the conditional independence relationships specified implicitly or explicitly by the GAPM. In a Graphical model the variables correspond to vertices in a graph, and influence relationships translate into edges. The graph will contain observable variables and unobservable variables (e.g. model parameters and variables that are too complicated to observe directly). The main advantages of the construction of a graphical model are: complexity of the model takes a manageable form, and the Graphical model can be used to derive and study the joint probability distribution in a systematic way.

The second question is answered in part by the fact that specific classes of graphical models (directed a-cyclical graphs) permit automatic generation of the joint probability distribution conditioned on the observed variables. For such cases computational techniques (MCMC methods) are available (as demonstrated in one of the examples), that allow efficient drawings from the joint distribution *regardless of analytical tractability*, and hence identification of the model parameters.

With a calibrated probabilistic model of the problem domain in hand, it is then 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.

## EXECUTIVE SUMMARY

---

This document is Deliverable D02.2 of the Fourth-Framework project OPUS. The OPUS project aims to develop and demonstrate .

### **Objectives of Work Package WP02**

The objective of this work package is :

- t.

The work in Work Package WP02 has resulted in one deliverable:

*Deliverable D02.1* : ???,

### **Objectives of Deliverable D02.2**

The objectives of this deliverable are:

This deliverable describes methodologies used first of all from a general project-wide viewpoint, thereby focusing on the different methodologies used within the OPUS project. The different methodologies will be compared and assessed. Thereafter, the individual tools will be presented (in the annexes).

### **Specific functions described in Deliverable D02.2**

The functions described in this deliverable are the following four main functions:

For each of these on-line functions several tools and methodologies are discussed and compared, the further development is described, results of preliminary testing and validation work are presented, and the integration of these tools within OPUS is discussed.

# 1. INTRODUCTION AND FRAMEWORK

---

## 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 in the public and private sector within Europe. The research is focused on developing an innovative methodology, incorporating statistical and database systems. Transport planning is a prominent example of a topic that uses multiple sources of data, and will be the main test case for OPUS, but the cross-sectoral nature of the research will be demonstrated through the inclusion of an application in the field of health information as another example.

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 will develop a general statistical framework for combining diverse data sources and specialise this framework to estimate indicators of mobility such as travel patterns over space and time for different groups of people. The project will undertake pilot and feasibility study applications in London, Zurich, Milan, and on a national level in Belgium. Methods for extending the framework to information aspects of the health domain will also be investigated.

The benefits of OPUS will be:

- 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 participants in the OPUS project are as follows:

### Research Organisations

- CTS (Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London), United Kingdom – Lead Partner
- DEPH (Department of Epidemiology and Public Health, Imperial College London), United Kingdom
- ETHZ (Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau), Switzerland

- FUNDP, Transport Research Group (Facultés Universitaires Notre-Dame de la Paix), Belgium

#### **Practitioners**

- Minnerva Ltd., United Kingdom.
- Survey and Statistical Computing, United Kingdom.
- Katalysis Ltd., United Kingdom.
- PTV AG, Germany
- Systematica, Italy.

#### **Public Bodies**

- Transport for London (TfL), United Kingdom.
- World Health Organisation (WHO), Italy.

## **1.2 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.

### 1.3 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.

### 1.4 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 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;
- The model may fill in observation gaps in the data or extend data into non-observable areas.

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.

## 1.5 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.

## 1.6 Relation with the OPUS life-cycle

Work Group 2:

Objectives

- To develop a generic statistical framework to enable the optimal combination of complex spatial and temporal data from survey and non-survey sources for the estimation of indicators of the socio-economic performance of urban and regional system. An important principle is that the framework will be generic and not tied to any specific application domain. It will however, take account of the particular characteristics of urban and regional systems.
- To ensure, in particular, that this framework can accommodate
  1. data sources that provide both direct and indirect information on the relevant population parameters
  2. data at different levels of aggregation
  3. the issues raised by the aging of sample survey data and the consequent need for updating
  4. the effect of sampling and non-sampling errors (including survey non-response and other sources of missing data) and
  5. the opportunities presented by new data streams from IST systems themselves.

## 1.7 Objective of the Deliverable

The objective of this deliverable is to present the outlines of a statistical framework for combining the information contained in separate datasets.

As far as possible, this will be based on successful applications of statistical methods and techniques for merging datasets in other areas in the literature.

The first task therefore is to conduct a quick (and by no means exhaustive) literature scan for successful solutions to problems similar to the one address in OPUS.

Given that solutions usually do not directly translate across applications domains, one of the first items for investigation will be a reflection on what allowed these methods to succeed and how best to apply successful methods developed in other problem domains to the specific problem addressed by the OPUS project. The aim of this reflection is to determine what would be transferable and how, and to compare that with the needs of OPUS.

Having taken into consideration what can be borrowed from other fields and what is needed within OPUS should allow us to assess what theoretical issues still need to be resolved, and what .

## **1.8 Structure of the deliverable**

A number of relevant problem instances will be presented in section 2. Section 3 delineates the way in which we propose to identify commonalities in the structure of problems that will enable us to apply methods and software developed in one problem instance to others. In section 4 we will list the formal methodologies that we have identified as most promising and appropriate for a generic formulation. In section 5 we will draw the observations together, and outline the work plan for the next step.

## 2. PROBLEM INSTANCES

---

### 2.1 Transport

#### 2.1.1 Transport modelling

Transport modelling can suitably be viewed as a means of calculating the interaction between transport demand and supply. Within this broad definition there are categories of models that vary by geographic scale (national, regional, urban, local, etc), by focus on travel modes (multi-modal, car, bus, rail, cycle, etc), as well as by focus on policies (demand management, urban traffic control, bus priority) and on schemes (new rail station, improved road, etc).

Transport modelling can also be divided into passenger and freight modelling. We follow the historic pattern of implicitly considering passenger travel (taken here to include car drivers), but OPUS should also be seen as offering a contribution to freight modelling.

Passenger models are classically organised into four stages (Ortuzar and Willumsen (2001)), namely, trip generation, distribution, mode choice, and assignment, though there is an increasing trend to introduce an initial stage of land use modelling that links with trip generation to result in a five-stage model. Relating to the view introduced above, all of these stages are part of demand modelling, except for assignment, which models supply, that is, the transport network and infrastructure.

The interaction between demand and supply raises problems of finding stable solutions. This is a matter of on-going attention, including from the UK Department of Transport's Variable Demand Modelling research [ref UK Department for Transport Seminar Papers on Variable Demand Modelling and DIADEM, 17<sup>th</sup> July 2003, Great Minster House, London].

From a transport policy and scheme assessment standpoint the derivation of stable solutions has great significant to aiding, rather than confusing, practical decision making. However, it may be noted that real life may not always correspond to 'stable solutions', especially in the presence of demand levels at or above network supply capacity levels. This complexity is normally ignored in transport modelling, but will become more apparent as automated sources of data become more available for analysis purposes, including exploitation by OPUS. The OPUS methodology will therefore need to be aware of the underlying stability of the data that it is using.

Transport demand models fall into two types of 'absolute' or 'incremental', depending on whether they forecast the absolute number of trips in future scenarios, or the incremental change from a base year. The latter, incremental, form of modelling is most robust with respect to errors and is therefore preferred when it is feasible to apply it. (It has limitations when forecasting aspects that are not present in the base year.)

The work of OPUS is especially relevant to the determination of accurate and rich representations of base year situations. 'Rich' in this context corresponds to the different dimensions in which transport data can apply and covers trip purpose, travel mode, time of travel, user characteristics, and so on. Transport modelling (as with other modelling) is strengthened by the availability of more disaggregate data, which an OPUS methodology should provide.

The extent of such data enrichment will be a key issue for transport modelling in respect of OPUS. One example is provided by two forms of demand survey data that are typically collected (and are particularly evidenced in the LATS 2001 transport survey data for London that will be used by OPUS Work Packages 4 and 8), namely, household data and traveller surveys. These both provide information on the traveller and the trips that are recorded, but they differ in that household data provides good samples of information on travellers but less good of their trips, while the reverse applies to traveller surveys. Linking this information

through OPUS methodologies would open the way for a more disaggregate style of transport modelling that would improve the sensitivity and accuracy of forecasts. As a simple example of the point, transport models typically ignore gender and age as explanatory variables, but these can be significant to many issues of transport choices.

Transport modelling is traditionally weak in the way it incorporates time as a dimension, although increased use of microsimulation models mean that there are more assignment models that can explicitly address time variations. Providing such models with time-varying demand information remains problematic and OPUS methodology will be of assistance in this regard.

While OPUS is naturally relevant to many of the concerns of transport demand modelling, it can also allow more detailed pictures to be obtained of transport system operating characteristics, for example, to show measures of congestion and accessibility. Such measures are typically obtained from modelling, as traditional surveys (e.g. ‘car following’) provide only patchy representations. However, increased control devices on the road network linked to ITC system means that there is a large amount of data that is under-exploited. Indeed, it is the volume of such data that makes it difficult to use. Also, while the data coverage is good, it does not provide a uniform coverage.

The OPUS approach, which blends data and modelling, is therefore well-placed to address this type of problem.

### **2.1.2 The use of prototypical samples in transportation demand models**

Several major transportation demand models in Europe (see Fox *et al.* (2003)) are based on the use of so-called prototypical samples.

The method of sample enumeration in to obtain population-level estimates of market share based on disaggregate discrete choice models is described in Ben-Akiva and Lerman (1984).

In large-scale transportation models this is often operationalised by conducting a detailed survey on a sample of the populations, which is selected to be as representative as possible for the population of the study area as a whole. Disaggregate choice models are then estimated on the sample.

In order to obtain population-wide predictions from these discrete choice models, the study area is divided into geographical zones, and the prototypical sample is re-weighted to match certain (disaggregate) totals (such as e.g. number of cars owned, number of people by age category, income class, household composition, etc.) for that zone, resulting in a set of zonal expansion factors. The discrete-choice models are then applied to the prototypical sample and re-weighted by the expansion factors to give the desired predictions at zonal level. By repeating this procedure for each geographical zone forecasts can be obtained at population level.

This procedure effectively combines a highly detailed dataset for a small subset of the total population, with a dataset containing with coarse, but population-wide totals.

The statistical reasoning that underpins this method is that the choice probabilities modelled depend exclusively on the attributes of the alternatives (as captured by the DCM), and are therefore *independent* of the zone, given the level of those attributes. This is equivalent to assuming that the behavioural discrete-choice models are geographically transferable.

### **2.1.3 Estimation of travel demand (mobility)**

As noted above, travel demand (mobility) is one of the two pillars of transport demand modelling, and its measurement is therefore of importance, and is a subject of a long history of research efforts. Unfortunately measurement of travel demand is not always easy.

Two central issues with the estimation of aspects of travel demand (mobility) are that:

1. the variables of interest are observed by a series of diverse measuring instruments, so that no single instrument reveals all aspects of travel demand
2. the different measuring instruments used all have their own have different statistical properties (e.g. with respect to variance, bias, level of observational, spatial and temporal detail, and cost)

The first issue makes it *inevitable* to use multiple data sources, and the second makes it *difficult*.

In Polak (2000a) we find a proposal for an estimation procedure using on a synthesis of existing data sources and domain-specific theory, which draws on the work of Ben-Akiva (1987), Ben-Akiva and Morikawa (1989), McNeil and Hendrickson (1985), and Polak (2000c). We have reproduced the proposal found in Polak (2000a) below.

### 2.1.3.1 General approach

The basic idea is to treat the problem of combining data from different sources as a problem of estimating the unknown population parameters of interest, taking explicit account of the specific characteristics of the data available from each source.

This is done by developing a model that expresses the probability of making certain measurements (by the different methods) conditional on the unknown population parameter values. With this model the likelihood (i.e., the joint probability) of the observed measurements having been made, conditional on the values of the parameters of interest can be calculated. The values of these parameters that maximise this likelihood give the best available possible estimates of the population parameters of interest.

This method proposed allows for the explicit characterisation of sampling errors, measurement biases and survey non-response, and can also accommodate methods that give indirect as well as direct estimates of the parameters of interest.

### 2.1.3.2 Natural Variation

Suppose that  $T_{ij}$  ( $1 \leq i \leq I, 1 \leq j \leq J$ ) is the mean number of trips in travel demand segment  $i$  made by persons in population segment  $j$ . The journey segments may correspond to different OD pairs or to different modes or to different purposes or to any combination thereof. Likewise, the population segments may be defined in terms of geographical location, working status, income etc. There are at total of  $I \times J$  such parameters and our aim to provide the best possible estimate of these parameters conditional on the available data.

There will be natural variation in the actual number of trips occurring during any given observation period (e.g. a day) due to a range of economic, social and seasonal factors. This source of variation must be properly taken into account when developing treatments of survey data.

In order to represent this ‘natural’ variation we assume that the number of trips actually being made on any randomly selected observation period is a random variable  $N_{ij}$ , which arises as the outcome of a Poisson process with mean  $T_{ij}$ . Thus,

$$N_{ij} \sim \text{Poisson}(T_{ij}) \quad (1)$$

that is,

$$\Pr(N_{ij}) = \frac{T_{ij}^{N_{ij}} e^{-T_{ij}}}{N_{ij}!} \quad (2)$$

Note that the assumption that  $N_{ij}$  is Poisson distributed is technically convenient (since it ensures that  $N_{ij}$  is always positive and integer valued, which is sensible for most travel demand measures) and parsimonious, but not crucial to the method. If other distributions are regarded as offering a better description of the natural variation in  $T_{ij}$  then these can be used instead.

We should noted in passing that natural (e.g., day-to-day) variation in travel demand can be an important source of apparent inconsistency between different surveys, if not properly accounted for in the data processing and consolidation phase. The magnitude of day to day variation in travel behaviour is surprisingly high. For example, research undertaken at Imperial College using 7-day travel diary data from the 1985/86 National Travel Survey showed that almost 45% of the total variation in daily trip rates was due to intrapersonal variation (Sanchez-Gomez, 1996). The implication of this is that even if the same data collection method was applied to the same sample of individuals but on different occasions we should expect to observe different results. Unless this additional source of variation is suitably parameterised using an appropriate statistical model it will be compounded with other sources of variation and will potentially distort subsequent analysis and inference with the data.

### 2.1.3.3 Measurement Bias

We assume that  $N_{ij}$  is measured by S alternative methods, giving rise to potential measurements  $Y_{ijs}$  for  $s = 1, \dots, S$ . The different methods may consist of different types of surveys (e.g., household travel diary surveys and in-vehicle surveys), census sources or model outputs. We assume that each measurement method is susceptible to method-specific measurement error. This means that although each method is measuring the same quantity, the measured values themselves will differ. It is reasonable to hypothesise that these measurement errors consist of two parts: (i) a systematic component or bias (e.g., in the case of travel diary surveys, under-reporting) which will probably be related in some way to the magnitude of  $N_{ij}$  and (ii) a random component due to sampling and related errors. It is probably reasonable to assume that this random component is normally distributed (although, as above this not crucial and other distributional assumptions are perfectly possible). Thus,

$$Y_{ijs} \sim \text{Normal}((N_{ij} + \beta_s(N_{ij})), \sigma_s^2) \quad (3)$$

that is,

$$\Pr(Y_{ijs} | N_{ij}) = \frac{1}{\sigma_s \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{Y_{ijs} - N_{ij} - \beta_s(N_{ij})}{\sigma_s} \right)^2} \quad (4)$$

where  $\beta_s(N_{ij})$  is a function giving the method-specific systematic measurement bias associated with method s and  $\sigma_s^2$  is the variance of the method-specific random measurement error (which we have assumed is constant across all ij pairs and with respect to  $N_{ij}$  - these assumptions could also be relaxed if necessary). Two special cases are of particular interest. If  $\beta_s(N_{ij}) \equiv \beta_s$  then the systematic error (bias) is additive, which might correspond for example to the effect of certain types of equipment and/or survey procedure failures. The second special case of interest occurs when  $\beta_s(N_{ij}) \equiv \beta_s N_{ij}$ , in which case the bias is multiplicative, as might be the case if a respondent in a travel diary survey persistently forgets to record details of e.g., return trips, short trips, non-motorised trips etc.

#### 2.1.3.4 Indirect Measurements

In the above we have assumed that the measurement methods each give direct estimates of  $N_{ij}$ . However, it is also possible to consider situations in which measurements are made by indirect methods, such as when link traffic counts are used to estimate origin-destination movements or station exit and entry counts used to estimate modal trip rates. In this case the relationship between the underlying measured quantity  $N_{ij}$  and the measurements actually made  $Y_{ijs}$  is more complex. For example, suppose that  $T_{ij}$  (and hence  $N_{ij}$ ) is the number of car trips between OD pair  $i$  by traveller group  $j$ . Further suppose that we make observations of the flow  $v_a$  ( $1 \leq a \leq A$ ) on a set of links.

If the measurements of flow are made without error then the relationship between the flow measurements  $v_a$  and the trip matrix  $N_{ij}$  will as follows:

$$v_a = \sum_{i,j,k} q_k^{ij} N_{ij} \delta(i, j, k; a) \quad (5)$$

where  $q_k^{ij}$  is the probability that travellers in class  $j$  will use the  $k$ th path between OD pair  $i$  and  $\delta(i, j, k; a) = 1$  if link  $a$  is on the  $k$ th path between OD pair  $i$  and is zero otherwise. The term  $\delta(i, j, k; a)$  captures the topological structure of the network and the choice probabilities  $q_k^{ij}$  capture the (route) choice process of travellers in the network. Typically  $q_k^{ij}$  are not observed directly but are the outcome of a modelling process. Essentially, equation (5) captures structural relationship between the measured ( $v_a$ ) and underlying ( $N_{ij}$ ) quantities, conditional on a (transport) model.

Indirect measurements will, of course, in general be subject to errors of the sort discussed in section 3.3, so in the instance of links flow, the measured flows be drawn from a distribution, so that

$$v_a \sim \text{Normal} \left( \left( \sum_{i,j,k} q_k^{ij} N_{ij} \delta(i, j, k; a) \right) + \beta(N_{ij}), \sigma_s^2 \right) \quad (6)$$

where we have again assumed that the random errors of observation are normally distributed.

More generally, with indirect measurements, we will be able to express the deterministic component of measured quantity as some function of the underlying parameters of interest,

$$Y_{as} \sim \text{Normal}((G_{ijas}(N_{ij}, \theta_s) + \beta_s(N_{ij})), \sigma_s^2) \quad (7)$$

where the function  $G_{ijas}(N_{ij}, \theta_s)$  captures the structural relationship between the measured quantity  $Y_{as}$  and the underlying quantity  $N_{ij}$  and  $\theta_s$  are a set method-specific structural parameters. As discussed above, in the case of traffic counts providing indirect information on OD demands, the term  $G_{ijas}(N_{ij}, \theta_s)$  would give the share of the  $N_{ij}$  trips intercepted at the counting station  $s$ . More generally,  $G_{ijas}(N_{ij}, \theta_s)$  can be thought of as a transport model that captures the relationships amongst structural quantities. Note that sometimes the  $G_{ijas}(N_{ij}, \theta_s)$  will be available in closed algebraic form (e.g., a logit mode choice model), whilst in others it will only be available implicitly via the operation of a wider modelling system such as LTS. The role of transport models in the context of data consolidation tasks will be discussed further in section 4.

### 2.1.3.5 Non-response Effects

With surveys that necessarily entail direct contact with and depend upon securing the cooperation of the travelling public, experience suggests that it is necessary to give serious consideration to the incident and implications of survey non-response. We can extend our framework in a relatively straightforward manner to accommodate this phenomenon.

Suppose for the sake of simplicity that all the individuals in the  $ij$  segment have the same probability  $p_{ijs}$  of responding to an attempt at contact via survey method  $s$ . The response probabilities  $p_{ijs}$  may either be derived empirically or (a better option) estimated on the basis of a model of non-response behaviour. For example, a simple logit model of the following form would suffice:

$$p_{ijs} = \frac{1}{1 + e^{\phi'_s Z_{ij}}} \quad (8)$$

where  $Z_{ij}$  is a vector of variables describing relevant characteristics of the  $ij$  segment (e.g., travel distance or travel time) and  $\phi_s$  is a method-specific parameter vector. One advantage of this model-based approach is that it reduces significantly the number of parameters that need to be estimated from the data (from potentially  $I \times J \times S$  to a figure of the order to  $S$ ).

If a particular attempted contact results in a successful survey then the measurement  $Y_{ijs}$  is made, otherwise no measurement is made. Under this assumption, we can define the final  $X_{ijs}$  as the outcome of a Bernoulli trial,

$$X_{ijs} \sim \text{Binomial}(Y_{ijs}, p_{ijs}) \quad (9)$$

that is,

$$\Pr(X_{ijs} | Y_{ijs}) = \binom{Y_{ijs}}{X_{ijs}} (p_{ijs})^{X_{ijs}} (1 - p_{ijs})^{Y_{ijs} - X_{ijs}} \quad (10)$$

### 2.1.3.6 Formulation of the Likelihood Function

We are now in a position to formulate the key relationship within our framework. For simplicity we initially concentrate on the case in which all the methods in question render direct estimates of the parameters of interest. Under this assumption, using equations (8), (10), (12) and (14) we can derive an expression for the probability of making the measurement  $X_{ijs}$  conditional on the true value of travel demand being  $T_{ij}$ . This probability is the product of the probability of actually making the measurement  $X_{ijs}$  conditional on their being a measurable value of  $Y_{ijs}$ , times the probability of a measurable value  $Y_{ijs}$  conditional on a true value of  $N_{ij}$  pertaining during the survey period, times the probability that a true value of  $N_{ij}$  pertains in the survey period conditional on the true average value being  $T_{ij}$ . Expressed algebraically this becomes:

$$\Pr(X_{ijs} | T_{ij}) = \underbrace{\Pr(X_{ijs} | Y_{ijs})}_{\text{non-response}} \underbrace{\Pr(Y_{ijs} | N_{ij})}_{\text{measurement}} \underbrace{\Pr(N_{ij} | T_{ij})}_{\text{natural variation}} \quad (11)$$

The functional form of the probability  $\Pr(X_{ijs} | T_{ij})$  will in general be rather complex, but modern methods of numerical simulation (e.g., Markov chain Monte Carlo methods – see e.g., Gilkes et al., 1996) make it possible to work with such complex distributions with relative ease. Closed form expressions for  $\Pr(X_{ijs} | T_{ij})$  can be obtained by judicious choice of the component distributions, if this is deemed desirable.

We can now express the likelihood of the set of measured values  $\{X_{ijs}\}$  conditional on the parameters of travel demand  $T_{ij}$ . The likelihood will also depend on the measurement bias and error parameters  $\beta_s$  and  $\sigma_s$  and on the non-response parameters  $\varphi_s$ . This likelihood has the form:

$$L(T_{ij}, \beta_s, \sigma_s, \varphi_s) = \prod_{i,j,s} \Pr(X_{ijs} | T_{ij}) \quad (12)$$

hence the Log-likelihood is:

$$\text{Log}L(T_{ij}, \beta_s, \sigma_s, \varphi_s) = \sum_{ijs} \text{Log}(\Pr(X_{ijs} | T_{ij})) \quad (13)$$

Maximising the value of the function  $\text{Log}L(T_{ij}, \beta_s, \sigma_s, \varphi_s)$  with respect to the parameters  $T_{ij}$ ,  $\beta_s$ ,  $\sigma_s$  and  $\varphi_s$  yields the maximum likelihood estimates of these parameters and the second derivative of the likelihood function yields information about the variances and covariances of the estimators. In this way, the proposed method is capable of rendering estimates not only of the substantive mobility parameters but also of important methodological parameters relating to survey biases and non-response behaviour.

Note finally that the likelihood presented in equations (18) and (19) above has been framed in terms direct observations. If a set of indirect observations are also available, these can be integrated into the analysis in a very straightforward manner, simply by extending the definition of the likelihood function as follows,

$$L(\text{Total}) = L(\text{Direct Measurements}) \times L(\text{Indirect Measurements}) \quad (14)$$

or equivalently,

$$\text{Log}L(\text{Total}) = \text{Log}L(\text{Direct Measurements}) + \text{Log}L(\text{Indirect Measurements}) \quad (15)$$

### 2.1.3.7 Scope of application

Although the proposed method was initially motivated by specific issues concerning the reconciliation of differences between a household survey and several on-board surveys, it is clear that it has much wider applicability to general problems of data consolidation in transport. In this section we briefly highlight a number of these potential applications.

#### Tracking changes in travel behaviour

Consider the issue of tracking through time changes in various aspects of travel demand, especially mode use and market shares. Monitoring implies relatively frequent surveys and therefore surveys that are also relatively small in terms of sample size and coverage. Although such surveys may be able to provide reasonable estimates of highly aggregate quantities, they are most unlikely to be able to provide credible estimates at a more disaggregate level. Yet many technical, political and policy questions are crucially dependent upon spatially and/or socio-economically ‘local’ effects.

The methods described in this paper would enable the development of systems which allow data from frequent, small ‘monitoring’ surveys to be used to consistently update disaggregate estimates derived from larger, less frequent surveys. Moreover, such a system could also, in principle, be extended to make use of real-time and quasi real-time data sources such as traffic counts and station exit and entry counts etc. Such a system would provide a means of bring to bear a degree of coherence on the way in information on travel demand, without having to impose in a Draconian fashion a single survey methodology.

The methods are also directly relevant to handling the data generated by a rolling survey design, where the need arises to optimally combine data returned in different ‘waves’ of survey work.

### **Integrating data from household surveys and on-board surveys**

A key issue at present is how best to reconcile data from household surveys and on-board surveys. The issue of deriving estimates of total household trip making from individual person sample is considered in more detail in Polak (2000b).

However, it is relevant to point out that whatever procedure is adopted to achieve this objective, the result will be alternative estimates of household mobility, which can effectively be considered to have been derived from different measurement methods. Hence, the procedure described above can be used to combine these separate estimators to produce a superior, pooled estimator.

### **Integration of Modelling Outputs and Survey Data**

The section 2.1.3.4 we briefly described how the methods outlined in this paper can be used to integrate indirect as well as direct measurements of parameters of interest. We also pointed out that in many cases, the use of indirect measurements (such as volumetric counts) will imply the need to also integrate certain model-based outputs (such as route choice proportions) with estimates from survey data.

It is important to appreciate however that the proposed methods offer a consistent framework for a far broader integration of models and model outputs with survey data and that it therefore has implications that extend well beyond the tasks associated with the estimation of a single set of travel demand parameters.

Existing transport models such as LTS and various local area traffic management models, play a potentially important role in consolidating data from different sources because they are a vital (indeed, the vital) source of both explicit and implicit structural information regarding the relationship between different measured data items. This was nicely illustrated by the example in section 2.1.3.4, in which link flows and OD demands were related to one another via network topology and route (modelled) choice behaviour. However, the range of application of such joint use of modelled and measured information is far broader than just OD matrix estimation.

We can effectively regard a transport model as offering an estimate of the first (and sometimes second) moment of the joint distribution of the spatially and temporally specific patterns of travel demands (e.g. OD demands), network flows (e.g., link flows, vehicle loadings) and level of service measures (e.g., travel time, fares etc.). Measurements of travel demand derived from different survey methods essentially provide estimates of specific marginal and conditional distributions (e.g., total number of trips, trips by specific modes etc.).

The methods described above provide an entirely general framework for combining these different sources of information to produce the best possible estimate of the underlying joint distribution. Such a joint distribution can be envisaged as a ‘database’ of (the best possible estimates of) demands, flows and levels of service, together with their associated standard errors. If such a database can be established through the judicious combination of multiple

data sources, it will provide a generic resource that can be queried to provide reports on a wide transport demand related topics. Critically, the existence of such a database would substantially reduce (perhaps even entirely obviate) the need for a succession of ad-hoc data consolidation exercises to be undertaken as different analysis issues arise.

### 2.1.4 Combined SP-RP estimation

In transport and in marketing two types of survey data are usually distinguished: SP (Stated Preference) data, and RP (Revealed Preference) data.

SP surveys basically ask a respondent for the preferred alternative in a hypothetical choice situation (“what would you do if ...?”), whereas RP surveys ask for the preferred alternative in an existing choice situation (“what did you do when ...?”).

Both types of survey have advantages and disadvantages. RP data is more firmly tied to reality as it describes real choices made by real individuals who had to choose under real constraints of time, information, budget, etc. RP data is often seen reasonable for short-term forecasts of small departures from the current state of affairs.

SP data on the other hand asks for a respondent’s preference under hypothetical conditions, and is therefore not limited to the situation as it exists at the time of the survey. It is therefore possible to more fully explore trade-offs and is often the only possible way to gain information about e.g. the market share of an alternative that does not exist at the time of the survey (e.g. a new travel mode or a new route alternative). Unfortunately SP data are more hypothetical than RP data.

It turns out that it is possible to combine these two sorts of data in a single model estimation, often known as combined RP-SP model estimation.

The idea is that a discrete-choice model (see Ben-Akiva and Lerman (1984) for a clear and thorough introduction) is estimated on the combined datasets. Assuming for a moment that the model is of the Multinomial Logit type, we can state the choice probabilities for a particular alternative  $i$  under RP and SP data as (see also Louviere *et al.* (2000):

$$P_i^{RP} = \frac{\exp[\lambda^{RP} (\alpha_i^{RP} + \beta^{RP} X_i^{RP} + \omega Z_i)]}{\sum_{j \in C^{RP}} \lambda^{RP} (\alpha_j^{RP} + \beta^{RP} X_j^{RP} + \omega Z_j)} \quad \forall i \in C^{RP} \quad (16)$$

$$P_i^{SP} = \frac{\exp[\lambda^{SP} (\alpha_i^{SP} + \beta^{SP} X_i^{SP} + \delta W_i)]}{\sum_{j \in C^{SP}} \lambda^{SP} (\alpha_j^{SP} + \beta^{SP} X_j^{SP} + \delta W_j)} \quad \forall i \in C^{SP} \quad (17)$$

with:

$C^{RP}, C^{SP}$	the choice sets for in the RP and the SP experiment respectively,
$P_i^{RP}, P_i^{SP}$	choice probabilities for alternative $i$ in the RP and the SP experiment
$X_i^{RP}, X_i^{SP}$	observable attributes that the RP and SP alternatives have in common
$Z_i, W_i$	the observable attributes unique to the RP and SP alternatives
$\alpha_i^{SP}, \beta^{SP}, \delta$	the model parameters for the SP situation
$\alpha_i^{RP}, \beta^{RP}, \omega$	the model parameters for the RP situation
$\lambda^{RP}, \lambda^{SP}$	the model scale parameters (made explicit for this occasion)

The idea is to combine the choice probabilities given by eqns. (16) and (17) into a single likelihood function for the combined dataset, by imposing the restriction that:

$$\beta^{SP} = \beta^{RP} = \beta \quad (18)$$

As it is not possible (see Louviere *et al.* (2000)) to identify both scale factors,  $\lambda^{RP}$  is usually set to 1. Collecting all model parameters into a single parameter vector:

$\psi = (\alpha^{RP}, \beta, \omega, \alpha^{SP}, \delta, \lambda^{SP})$  gives the following combined likelihood function:

$$L(\psi) = \sum_{n \in RP} \sum_i y_{in} \ln P_{in}^{RP}(X_{in}^{RP}, Z_{in} | \alpha^{RP}, \beta, \omega) + \sum_{n \in SP} \sum_i y_{in} \ln P_{in}^{SP}(X_{in}^{SP}, W_{in} | \alpha^{RP}, \beta, \delta, \lambda^{SP}) \quad (19)$$

By using the likelihood function in eqn. (19) to identify the parameter vector  $\psi$ , an estimate of all the parameters in eqns. (16) and (17) is obtained, thereby *effectively merging the information of the RP and SP datasets*.

Examples of the application of this technique can be found in Bradley and Kroes (1990), Bradley and Daly (1992), Ben-Akiva and Morikawa (1990) and see Louviere *et al.* (2000).

We note that the critical part is the formulation of the joint likelihood function in eqn. (19). Once this is accomplished, (provided that the model specified by the joint likelihood is actually correct, or “good enough”) the actual estimation of the parameters on the joint dataset (which could be interpreted as the actual merging of the datasets) is a question of technique.

## 2.1.5 Positioning

As an example of the application of the combination of different data sources we refer to the paper by Zhao *et al.* (2003), which uses Kalman filtering to determine the position of a moving vehicle in real-time by filtering the noise from observational data and by integrating three different data sources. In order to illustrate the process, we have partly reproduced this article below.

### 2.1.5.1 Background

The objective is to provide continuous and accurate positioning information for a driving vehicle. In this vehicle positioning we have two sources of position information: a Dead-Reckoning (DR) system consisting of an odometer and a gyroscope, and a satellite navigation system (GPS), both of which have strengths and weaknesses.

The DR supplies virtually noiseless outputs that slowly drift off with time, whereas the GPS has minimal drift but much more noise and may be locally degraded (e.g. in urban areas where the satellite signal is reflected by buildings) or even unavailable (e.g. in tunnels).

Based on existing theory, a dynamic (continuous-time) model was formulated of how the vehicle is expected to move. Based on the (known) error structures of the DR system and the GPS system, observation equations were drawn up.

These equations were then discretised and linearised around the true trajectory of the vehicle so that a Kalman filter could be set up to integrate the sources of information into a single position estimate that combines the minimal drift of the GPS with the accuracy and high availability of the DR system. An understanding of the error structure of both DR and GPS is crucial.

### 2.1.5.2 The transition equations

The following states were selected:

$$\mathbf{x} = [e \quad n \quad v_v \quad H_v \quad a \quad \omega \quad \delta S \quad \delta K \quad \varepsilon_G]^T \quad (20)$$

Where

- e = terrestrial easting position, in meters
- n = terrestrial northing position, in meters

- $v_v$  = forward velocity of the vehicle, in m/s, with forward being positive  
 $H_v$  = heading of the vehicle, in radian, with north being zero and clockwise being positive  
 $a$  = the acceleration of the vehicle, in  $m/s^2$   
 $\omega$  = the rate of the heading, in rad/s  
 $\delta S$  = the odometer scale factor error, in m/pulse  
 $\delta K$  = the rate gyro scale factor error, in mv/rad/s  
 $\varepsilon_G$  = the bias drift of the rate gyro, in rad/s

The (continuous-time) dynamic equations can be written as:

$$\left. \begin{aligned}
 \dot{e} &= v_v \cdot \sin H_v + w_1 \\
 \dot{n} &= v_v \cdot \cos H_v + w_2 \\
 \dot{v}_v &= a + w_3 \\
 \dot{H}_v &= \omega + w_4 \\
 \dot{a} &= w_5 \\
 \dot{\omega} &= -\beta_\omega \omega + w_6 \\
 \dot{\delta S} &= w_7 \\
 \dot{\delta K} &= w_8 \\
 \dot{\varepsilon}_G &= -\beta_g \varepsilon_G + w_9
 \end{aligned} \right\} \quad (21)$$

Or in vector form:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}(t), t) + \mathbf{w} \quad (22)$$

Where  $\mathbf{w} = [w_1 \ w_2 \ w_3 \ w_4 \ w_5 \ w_6 \ w_7 \ w_8 \ w_9]^T$  is the dynamic noise. The variables  $\beta_w, \beta_g$  are the skew correlation times (i.e. inverse of correlation times). It can be seen that this is a non-linear set of equations because there are two items that are non-linear.

### 2.1.5.3 The observation equations

From the GPS receiver, the information on position ( $\varphi_{GPS}, \lambda_{GPS}$ ), velocity  $v_{GPS}$  and heading  $H_{GPS}$  of the vehicles can be derived. From the odometer and rate gyroscope, the pulses  $\Delta N_{odo}$  during a time interval  $\Delta t$ , representing the displacement travelled, and direct current (DC) voltage output  $V_{RG}$ , representing the heading-rate of the vehicle respectively, can be acquired. So the observation variables are chosen as follows:

$$\mathbf{z} = [\lambda_{GPS} \ \varphi_{GPS} \ v_{GPS} \ H_{GPS} \ \Delta N_{odo} \ V_{RG}]^T \quad (23)$$

And the measurement equations are:

$$\left. \begin{aligned}
 \lambda_{GPS} &= e/R \cdot \cos \varphi_{GPS} + v_2 \\
 \varphi_{GPS} &= n/R + v_1 \\
 v_{GPS} &= v_v + v_3 \\
 H_{GPS} &= H_v + v_4 \\
 S \cdot \Delta N_{odo} &= 1/2 \cdot a \cdot \Delta t^2 + v_v \cdot \Delta t - \delta S \cdot \Delta N_{odo} + v_5 \\
 V_{RG} &= (K + \delta K)(\omega + \varepsilon_G) + v_6
 \end{aligned} \right\} \quad (24)$$

Where  $S$  and  $K$  are the nominal scale factor for the odometer and gyroscope respectively;  $\Delta t$  the interval time;  $R$  the radius of the earth. These equations can be rewritten in matrix form as:

$$\mathbf{z} = \mathbf{h}(\mathbf{x}(t), t) + \mathbf{v}$$

(25)

Where  $\mathbf{v} = [v_1 \ v_2 \ v_3 \ v_4 \ v_5 \ v_6]^T$  is the observation noise. It can be seen that the measurement equations are also non-linear because of the last measurement equation.

#### 2.1.5.4 Kalman Filter Design

The Kalman filter can be used to produce optimal estimates of the state vectors listed above with well-defined statistical properties. For convenience of computer calculation, discrete recursive algorithms are usually adopted. So the first thing to do is to discretise the continuous dynamic equations. In addition, due to the non-linear properties of both the system dynamic equations and measurement equations, linearisation is also required. The state transition and observation matrices are derived as follows (Gelb, 1979):

$$\left. \begin{aligned} \mathbf{x}(k) &= \Phi(k, k-1)\mathbf{x}(k-1) + \mathbf{w}(k-1) \\ \mathbf{z}(k) &= \mathbf{h}(\mathbf{x}(k), k) + \mathbf{v}(k) \end{aligned} \right\} \quad (26)$$

Where

$$\Phi(k, k-1) = \mathbf{I} + \left. \frac{\partial \mathbf{f}(\mathbf{x}(t), t)}{\partial \mathbf{x}(t)} \right|_{\substack{\mathbf{x}(t)=\mathbf{x}(k) \\ \mathbf{w}(t)=0}} \cdot \Delta t$$

$$= \begin{bmatrix} 1 & 0 & \sin H_v \cdot \Delta t & v_v \cdot \cos H_v \cdot \Delta t & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & \cos H_v \cdot \Delta t & -v_v \cdot \sin H_v \cdot \Delta t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \Delta t & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 - \beta_\omega \cdot \Delta t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 - \beta_g \cdot \Delta t \end{bmatrix}$$

$$\mathbf{H}(k) = \left. \frac{\partial \mathbf{h}(\mathbf{x}(t), t)}{\partial \mathbf{x}(t)} \right|_{\substack{\mathbf{x}(t)=\mathbf{x}(k) \\ \mathbf{v}(t)=0}}$$

$$= \begin{bmatrix} 1/R \cos(\varphi_{GPS}) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/R & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \Delta t & 0 & \Delta t^2/2 & 0 & -N_{odo} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & K + \delta K & 0 & \omega + \varepsilon_G & K + \delta K \end{bmatrix}$$

$$\text{cov}(w_k, w_j^T) = Q_k \delta_{kj};$$

$$\text{cov}(v_k, v_j^T) = R_k \delta_{kj};$$

$\delta_{kj}$  is the Kronecker- $\delta$  function;  
 $\Delta t$  is the sampling time interval;  
 $k, j$  are the discrete points in time;

The result was an uninterrupted series of position estimates with the low drift of the GPS system and the high accuracy of the Dead-Reckoning system

## 2.2 Health

An example of the application of the type of methodology that the OPUS project envisages in an epidemiological context can be found in Richardson (1996), which we have partially reproduced here.

The example uses three models:

- A “disease model” that links the disease status  $Y_i$  of individual  $i$  given directly observable risk factors  $C_i$ , not always directly observable risk factors  $X_i$ , and a parameter  $\beta$ .
- A “measurement model” for the relationship between observable proxy (or surrogate) variables  $Z_i$  for the (not always observable) risk factors  $X_i$ , the risk factors  $X_i$  themselves, and a parameter  $\lambda$
- An “exposure model” for the relationship between the not always observable risk factors  $X_i$ , the observable risk factors  $C_i$  and a parameter  $\pi$

Adopting the notation of the article, where  $[a]$  denotes the probability distribution of  $a$ , and  $[a | b]$  the probability distribution of  $a$  given  $b$ , we can denote the model equations as:

$$[Y_i | X_i, C_i, \beta] \quad (27)$$

$$[Z_i | X_i, \lambda] \quad (28)$$

$$[X_i | C_i, \pi] \quad (29)$$

The article then notes that adoption of eqns. (27)-(29) implies that the joint probability distribution of the stochastic quantities  $X_i, Y_i, Z_i, \beta, \lambda, \pi$  can be written as:

$$[X_i, Y_i, Z_i, \beta, \lambda, \pi] = [\beta][\lambda][\pi] \prod_i [X_i | C_i, \pi][Z_i | X_i, \lambda][Y_i | X_i, C_i, \beta] \quad (30)$$

In particular:

- (27) implies that  $Y_i \perp Z_i | X_i, C_i, \beta$ , i.e.  $Y_i$  is independent of  $Z_i$  given  $X_i, C_i$  and  $\beta$
- (28) states that  $Z_i \perp Z_j | X_i, \lambda \quad \forall i \neq j$ , i.e. the proxies  $Z_i$  for individuals are mutually independent given the true exposures  $X_i$  and the parameters  $\lambda$
- (29) states that the risk factors  $X_i$  for individuals are mutually independent given  $C_i$  and  $\pi$

The relationship between the variables specified in eqns. (27)-(29) can be depicted graphically as in Figure 1. Observed variables are shown in squares, and unobserved variables are shown in ovals.

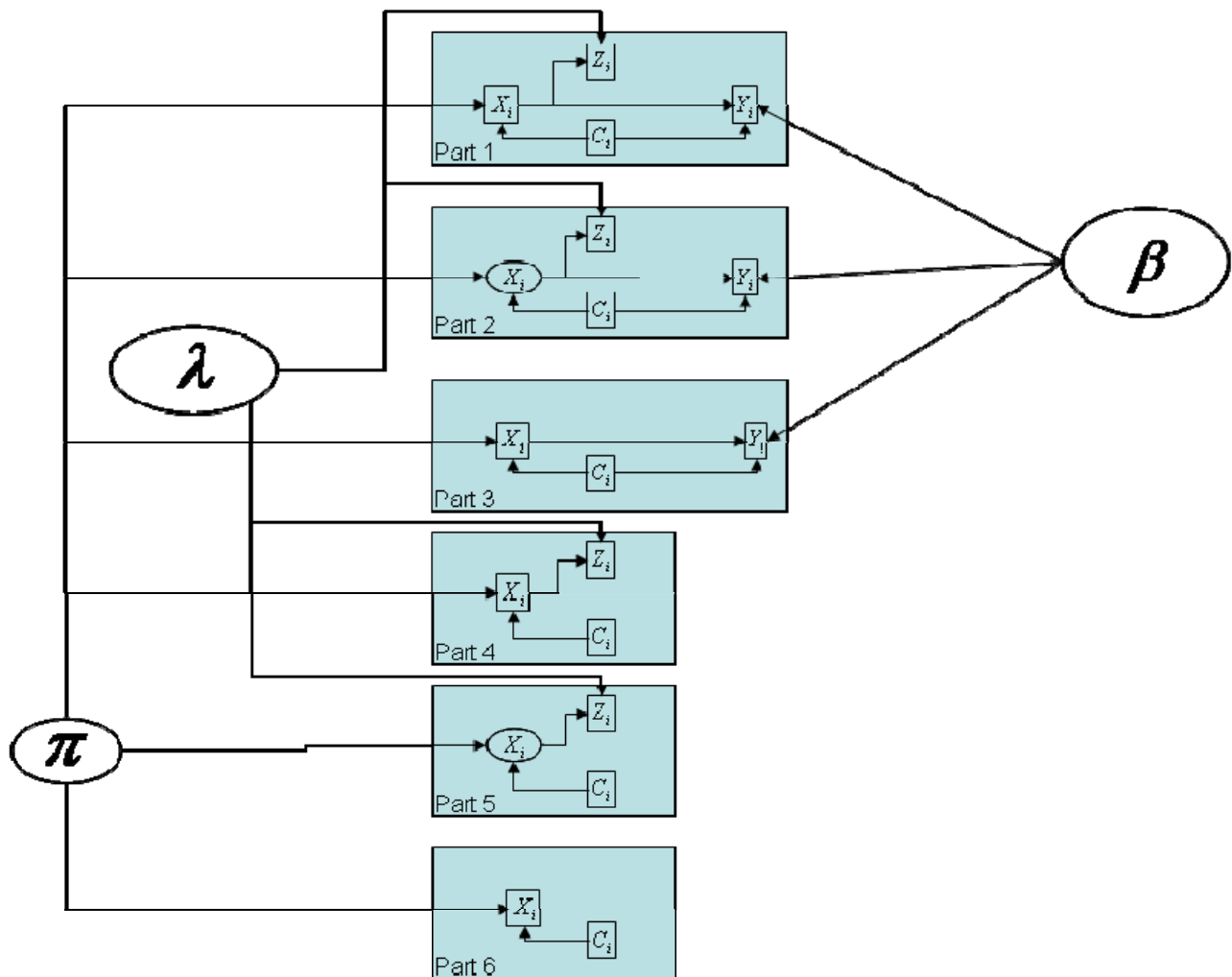


Figure 1: an influence graph for the models under various epidemiological designs

This figure shows 6 groups of individuals, according to what variables are observed. The shaded boxes correspond to different observational setups, and show the models and the variables observed. Each box has a part number, shown in the bottom left corner. The models will be described by part no.. Sometimes the exposure  $X_i$  can be observed without error; this is a rare and valuable event which serves as a “gold standard” for the model.

Part no.:

1. corresponds to a so-called internal validation study; all variables are observed, even those that are normally unobserved
2. corresponds to the common situation where only surrogates and the disease are known
3. represents a subgroup in which only the true exposure and disease status are known
4. corresponds to a so-called external validation study; all variables are observed, except the disease status
5. corresponds to a “survey” situation, in which information is obtained only on proxy variables
6. corresponds to a “survey” situation, in which information is obtained on the true exposure

The arrows in Figure 1 represent the influence of variables and parameters in the model.

Figure 1 also illustrates how information may flow from one part of the diagram to another. When observations are available on a variable that an arrow points to (e.g.  $X_i$  in part 6), then this information may be used to infer something about the variable(s) from which the arrow originates (e.g. the parameter  $\pi$ ). When information is available on  $\pi$ , observations on part 5 may be used to obtain information on parameter  $\lambda$ , etc.

When no “gold standard” (i.e. error-free measurement of  $X_i$ ) is available, estimation of the model parameters must rely on other sources of information, such as the use of different measuring instruments for the same variable, possibly with repeated measurements. In the notation of the original article, let  $Z_{ihr}$  denote the  $r^{th}$  repeated measurement of instrument no.  $h$  for individual  $i$ , then the measurement model in eqn. (28) becomes:

$$[Z_{ihr} | X_i, \lambda_h] \quad (31)$$

### 2.2.1 Example

We will now list one of the examples given in Richardson (1996), in which two measuring instruments were used.

Two risk factors for the disease were introduced:  $X$  and  $C$ , of which  $C$  is known accurately and  $X$  is subject to measurement error. We also assume that  $Y_i$  follows a Bernoulli distribution with parameter  $p_i$ , where:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i + \beta_2 C_i \quad (32)$$

It is furthermore assumed that:

$$(X, C) \stackrel{d}{=} N(\mu, \Sigma)$$

This means that  $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$  in eqn. (27) and that  $\pi = \begin{pmatrix} \mu \\ \Sigma \end{pmatrix}$  in eqn. (29).

The two measuring instruments have the following characteristics:

- Instrument 1 has a large variance (low precision), but is unbiased
- Instrument 2 has a small variance (high precision), but is known to be biased.

In view of its high precision, Instrument 2 is used on the whole population. However in order to be able to correct the bias of Instrument 2, both instruments are applied to a subsample of the total population; Instrument 1 being used twice on each individual.

The following model is assumed for the  $r$ 'th repeat of Instrument 1 is:

$$[Z_{i1r} | X_i, \theta_1] \stackrel{d}{=} N(X_i, \theta_1^{-1}) \quad r = 1, 2 \quad (33)$$

with  $\theta$  the precision (i.e. 1/variance), and  $\lambda_1 = \theta_1$ .

For Instrument 2, the model is:

$$[Z_{i2} | X_i, \phi_2, \psi_2, \theta_2] \stackrel{d}{=} N(\phi_2 + \psi_2 X_i, \theta_2^{-1}) \quad (34)$$

$$\text{and } \lambda_2 = \begin{pmatrix} \phi_2 \\ \psi_2 \\ \theta_2 \end{pmatrix}$$

### 2.2.2 Simulation study

A dataset of simulated observations was created using the values for  $\beta_0, \beta_1, \beta_2, \theta_1, \phi_2, \psi_2, \theta_2$  as shown in the columns “true values” of Table 1, and the following values for  $\mu$  and  $\Sigma$

parameter	true value	Gibbs sampling analysis				Classical analysis	
		N=200		N=50		With Instrument 2	
		Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
$\beta_0$	-0.8	-0.81	0.32	-0.77	0.40	-0.17	0.11
$\beta_1$	0.9	1.03	0.36	0.98	0.42	0.14	0.07
$\beta_2$	1.2	1.25	0.14	1.36	0.21	1.57	0.11
$\theta_1$	0.3	0.31	0.03	0.25	0.04	-	-
$\phi_2$	0.8	0.81	0.11	0.84	0.14	-	-
$\psi_2$	0.4	0.44	0.09	0.45	0.14	-	-
$\theta_2$	0.9	0.91	0.06	0.96	0.13	-	-

Table 1: Gibbs sampling analysis of a design with 2 measuring Instruments

$$\mu = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}, \Sigma = \begin{pmatrix} 1.02 & 0.56 \\ 0.56 & 0.96 \end{pmatrix}$$

This dataset was then analysed using Gibbs sampling analysis as described in Richardson and Gilks (1993), and using classical analysis using only the data from part 1, without correcting for measurement error, leading to the results shown in columns 3-6 of Table 1.

This example clearly shows the need to correct for measurement error, and illustrates the power of the Gibbs sampling analysis in doing so.

### 3. GENERAL APPROACH

In this chapter we will outline the general approach proposed by the OPUS project to the problems listed in chapter 2.

In order to do this, we will first (in section 3.1) propose a conceptual hierarchy which we use to order concepts such as the physical object system we are interested in, the identification of relevant concepts within the description of such a system, any theories we may have about it, their reflection in mathematical models, the statistical aspects associated with actually observing aspects of the system, the software used to operationalise the mathematical model used, and finally its application.

Next (in section 3.2) we will discuss the issue of finding commonalities between the problem instances that would allow us to develop generic models (and/or software) for those problems, and the limits imposed by the specifics of each problem.

In section 3.3 we will present a discussion of how a generic a priori model of a problem domain can be mapped onto mathematical and statistical models, and in section 3.4 we will outline the general approach proposed for the operationalisation of the OPUS vision.

#### 3.1 Conceptual hierarchy

In order to structure our discourse, we will use the conceptual hierarchy listed in Table 2 and shown in Figure 2 .

Nr	Level	Description
7	Application	
6	Algorithm	How do we calculate along the relationships
5	Statistical model	Where do the measurement errors come in?
4	Mathematical model	How do we model the influence relationships?
3	Theory	How do the variables influence each other, which ones can we observe?
2	Conceptual	What variables do we distinguish, and which do we want to know
1	Physical	What is the domain of discourse

*Table 2: The conceptual hierarchy adopted*

The hierarchy listed in table Table 2 is shown in Figure 2 as a sequence of selections and enlargements. Only part of the physical reality (shown at the bottom) is translated into the conceptual level. This translation can be carried out in various ways, i.e. there is considerable freedom regarding which concepts to define and use for a given part of the physical reality.

A subset of the concepts distinguished are tied together in a theory (of scientific or engineering quality), where again there is considerable freedom as to what form this theory should take.

This is repeated for the mapping of of theory onto a mathematical model, the extension of the of the mathematical model with an error structure, the mapping of the stistical model onto algorithms and operational software, and finally its application.

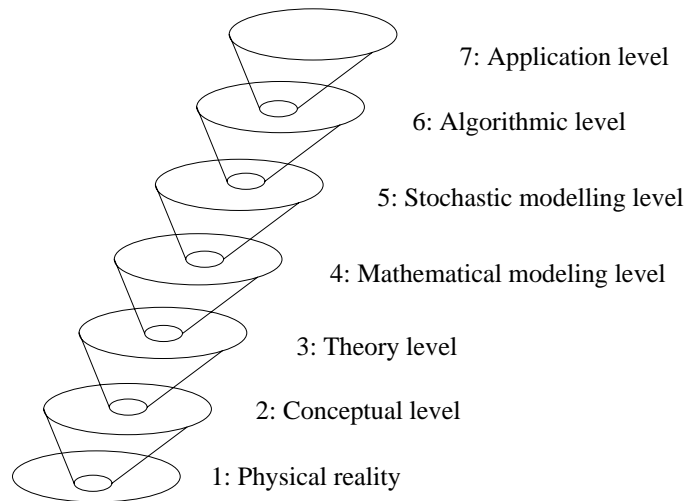


Figure 2: the conceptual hierarchy

### 3.1.1 An example: O-D matrix estimation

Consider the question of estimation of a table of trips between geographical zones (commonly known as the estimation of an Origin-Destination (OD) matrix).

Denote the number of trips from zone  $i$  to zone  $j$  as  $T_{ij}$ , the number of vehicles counted on arc  $a$  of the transport network as  $q_a$ , and the number of vehicles taking route  $p$  from  $i$  to  $j$  as  $f_{ij,p}$ .

Then at the physical level there would be the trips made by users of the transport system from their starting address to their destination address. Conceptually (on level 2) we often divide the geographical region into zones, and simply consider the number of trips from zone  $i$  to zone  $j$  as the conceptual unit of choice. Based on our understanding of the transport system, we have a theoretical model in which the trip table  $T_{ij}$  gives rise to route flows  $f_{ij,p}$ , which in turn give rise to counts on certain links on level 3.

This relationship can be formalised in a variety of mathematical models, of which we use the so-called ‘Stochastic User Equilibrium’ (SUE) as an example. From a mathematical point of view we have a mapping, called SUE, which takes the O-D matrix  $T_{ij}$  and produces link counts  $q_a$ .

Observations of link counts contain an error component, denoted as  $\varepsilon_a$ , which result in perturbed counts  $\tilde{q}_a$  being observed. When viewed from this point of view, all that we will ever be able to know about the quantity of interest is a probability distribution of its values,  $\tilde{T}_{ij}$  which we can derive from our observations of  $\tilde{q}_a$ . The joint probability distribution of the errors and the O-D matrix estimate  $\tilde{T}_{ij}$  may be calculated from the observations of  $\tilde{q}_a$  and the structural relationships  $T_{ij} \xrightarrow{SUE} q_a$  using variants of the so-called Markov Chain Monte Carlo algorithm (see chapter 4.2.2.4).

The conceptual hierarchy applied to the question of O-D matrix estimation is shown in Table 3

Nr	Level	Description
7	Application	
6	Algorithm	MCMC, BUGS
5	Statistical model	$\tilde{q}_a = q_a + \varepsilon_a; \tilde{q}_a \rightarrow \tilde{T}_{ijj}$
4	Mathematical model	$T_{ij}^{SUE} \rightarrow q_a$
3	Theory	$T_{ij} \rightarrow f_{ij,p} \rightarrow q_a$
2	Conceptual	$T_{ij}$
1	Physical	trips

Table 3: The conceptual hierarchy applied to O-D matrix estimation

### 3.2 Finding commonalities and discrepancies

Each problem instance differs at levels 1-3. This means different variables and different relationships.

Commonalities occur at levels 4 and 6, and often at level 5; this means that commonalities can only be expected at a certain level of abstraction. Due to the close relationships between statistical estimators and the problem structure, it remains to be seen at what level of abstraction methods become transferable.

### 3.3 Generic methods and specific instances

As noted in Logie (2003), 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).

#### 3.3.1 Constructing a starting point

Whilst in principle one might decide to infer both the structure and the strength of the interaction between variables in a problem domain from observational data (known in the literature as *unsupervised learning* (see e.g. Duda *et al.*, (2001))), we feel this approach is best applied when no meaningful theory of the problem domain exists.

In the field of Transport by contrast, a substantial body of meaningful theory exists (see e.g. Ortuzar and Willumsen (2001), and Cascetta (2001)). This project will therefore focus on problem domains for which an adequate theoretical framework exists, from which generic a-priori models may be derived.

The methodology to fill-in ‘missing data’ can be represented as a form of optimisation problem. This makes it easier to appreciate the dictum of optimisation that the best way to find an optimum solution in non-trivial cases is to start the search as close to the optimum as possible. In other words, establishing a suitable starting point is an important element of the OPUS method.

The vision is that OPUS provides (or perhaps uses or assumes as the model is not exhaustive) a ‘generic a priori model’ (GAPMs) for the subject areas of Transport and one in the general area of Health. These GAPMs provide stable, holistic views that are widely applicable. It is the task of specific applications to adjust the details to particular interests.

#### 3.3.2 A transport General A Priori Model (GAPM)

The primary components of a generic GAPM for transport are illustrated in Figure 1, 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), and Cascetta (2001).

The GAPM represents this information through modelling. The initial GAPM should reflect the important issues in a logical and intuitive manner so that:

- Changes to either demand and supply affect the other with plausible elasticities
- Changes in one part of the system influence others to a plausible extent
- Constraints, whether physical or policy-related, limit the nature of solutions provided by the model.

It is not a requirement that the GAPM provides an accurate validation but that its behavioural characteristics are plausible. The amount of detail represented by the GAPM is to be determined in the OPUS project; it does not necessarily have to incorporate large amounts of detail.

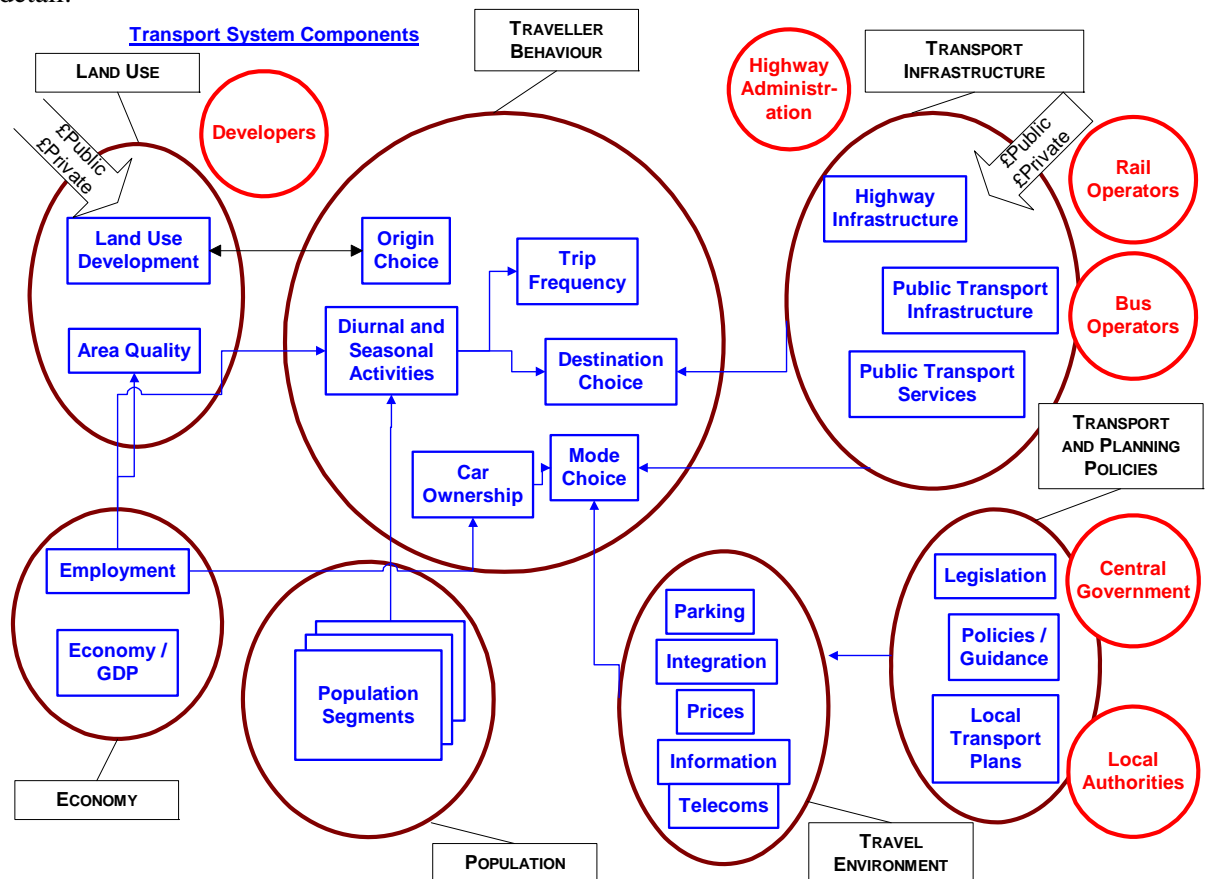


Figure 3: A GAPM for transport

### 3.3.3 Interpretation of a GAPM in terms of a graphical model

The GAPM proposed in section 3.3.2 covers the levels 2-3 as distinguished in section 3.1, by specifying:

- concepts that are commonly considered relevant for a transportation planning application

- their theoretical inter-relationships.

Note e.g. 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 is crucial, because it allows translation of Figure 3 into a statistical framework known as a graphical model (see Whittaker (1998)) in a relatively straightforward way. In fact, the requirement that the conditional independence between variables can be represented as a graph *defines* graphical models.

In doing so, the essential structure of the problem domain (in this case a transport model) may be captured in a statistical model.

### 3.4 General approach proposed

On basis of the considerations in this chapter, we propose the following general approach to operationalise the OPUS vision

- Formal description of variables and their structural relationships in terms of a graphical model
- Formal description of the link between observable quantities and structural variables; identification of the path in the graphical model between what is observed and what is not; problem-specific treatment of inference between any observable quantity and its nearest structural neighbours in the graph
- Construction of the joint a posteriori probability density of all variables in the graphical model conditional on the observations
- Extraction of relevant point estimates and moments from this a posteriori distribution

---

## 4. FORMAL METHODS AND SOFTWARE

---

As emerged from the discussion in chapter 3, the project will have to deal with the complexity models of a particular problem domain as represented by the GAPMs. An additional complication is that several variables that play a role in the GAPM are not directly observable, or leastways not without observation errors, which must be taken into account by any practical model. Finally, as we are in a situation where we require multiple datasources simultaneously, we will have to cope with datasources that contain missing observations.

Therefore formal methods are required that can cope with all three aspects of the problem:

- complex interrelationships between structural variables within the model
- errors that result from the observation process
- coping with missing observations.

### 4.1 Literature scan

In order to identify the most promising building blocks from which to build a modelling framework for the OPUS project, a quick literature scan was conducted.

First of all, the class of models called “Structural Equations Models” (SEM), as described in Bollen (1989), stands out. It is an established statistical framework that can handle complicated structure, is easily solved using linear algebra, distinguishes between observable and latent variables from the ground up, and separates structural equations from measurement equations. Unfortunately it has two drawbacks: it is restricted to *linear* models of the subject matter which seems quite inappropriate in the transport domain, and it is restricted to first and second moments.

A more recent development, known as Graphical Modelling as first described in Whittaker (1998), provides a much more general framework. It has all the capabilities of SEM models, but is not restricted to linear relationships, and it has no built-in limitations that restrict its use to first and second moments. For all its advantages, its use can be computationally intensive and until very recently would have been so far removed from practical implementation as to be mainly of theoretical interest.

Fortunately, the work of Spiegelhalter *et al.* (1999), Gilks *et al.* (1996), Spiegelhalter *et al.* (1996)), provides tools with which graphical models can be operationalised. As an extra benefit, the operationalisation follows a Bayesian approach, in which a full a-posteriori probability distribution is estimated rather than its moments.

Unrelated to this stream of work, an account of identification problems in the Social Sciences can be found in Manski (1995), and a systematic treatment of imperfectly observable variables is given in Manski (2003). This stream of work complements and extends the classical econometric tools of instrumental variables and structural equations as found in standard econometric textbooks, such as e.g. Amemiya (1985). As a matter of interest, in the Transportation domain the most important observational procedure to measure users’ preferences (the Stated-Preference survey; see e.g. Richardson *et al.* (1995)) sometimes requires complex weighing and correction procedures and is vulnerable to bias through non-response (see e.g. Bethlehem *et al.* (1986), and Bethlehem *et al.* (1987)). A related effect has been observed in panel surveys, where a method for removing bias has been proposed in Kitamura and Bovy (1987).

This has knock-on effects for certain crucial elements of the models used in transportation planning that are directly based on sample surveys: Discrete Choice Models (see e.g. Ben-Akiva and Lerman (1984)). The point is that the very GAPMs of the domain of discourse may contain specification errors, which should be taken into account in the OPUS approach.

A proposal for a model-based approach towards combining different datasets can be found in Logie (2001).

## 4.2 Representing the GAPMs of the problem domain

As noted in section 3.3.2, one of the essential features of the approach taken in OPUS is the provision of an *Generic A Priori Model* (GAPM) of the problem domain. Both Structural Equations Models and the more general Graphical models seem to be suitable vehicles to represent the structural relationships implicit in the GAPMs.

We will borrow from the terminology of SEMs and call the variables defined in the GAPM *structural variables*, which are often *latent* (i.e. not directly observable). When they are observable, they are called *manifest* but observational errors see to it that only a perturbed version of the variable in question will be observed, consisting of the true value and a random error term (which may or may not have zero mean).

### 4.2.1 Structural equations models

Structural equations modelling is described in various sources, of which we mention Bollen (1989)).

Structural equations models take the following form:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \quad (35)$$

$$\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \quad (36)$$

$$\mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (37)$$

Eqn (35) captures the *structural relationship* between latent variable vectors  $\boldsymbol{\eta}$  and  $\boldsymbol{\xi}$ ; where  $\mathbf{B}$  and  $\boldsymbol{\Gamma}$  are coefficient matrices. Eqns. (36) and (37) represent the *measurement equations* that relate the latent variables  $\boldsymbol{\eta}$  and  $\boldsymbol{\xi}$  to observable quantities  $\mathbf{x}$  and  $\mathbf{y}$ .

Structural equations models have the numerous advantages; SEMS:

- can handle complicated structures, as the vectors  $\boldsymbol{\eta}$  and  $\boldsymbol{\xi}$  and the corresponding coefficients of the structural equations  $\mathbf{B}$  and  $\boldsymbol{\Gamma}$  can have arbitrary dimension
- can be solved routinely using generic computer packages for linear algebra, or by dedicated software since eqns. (35) -(37) are all *linear*
- distinguish between observable and latent variables from the ground up
- separate structural equations from measurement equations

Unfortunately SEMs also have certain disadvantages: in that they are restricted to

- *linear* models
- first and second moments

On balance it was felt that SEMS, whilst very powerful in themselves, and quite practical in terms of their solution, fall short of the requirements of OPUS.

### 4.2.2 Graphical models

Graphical Models, as described in Whittaker (1998) encompass the SEM models, and extend them in three dimensions:

1. by relinquishing the distributional assumption that all stochastic components are normally distributed
2. by loosening the requirements of linearity in the relationships between variables

3. by relinquishing the parametric framework and allowing a Bayesian approach to be used

Doing this allows the modeller to focus on the more fundamental property of conditional independence as implied by the structure of the problem domain, as for example incorporated in any GAPMs that may be available.

The dimensions along which Graphical Modelling extends on Structural Equations Modelling are shown in Figure 4.

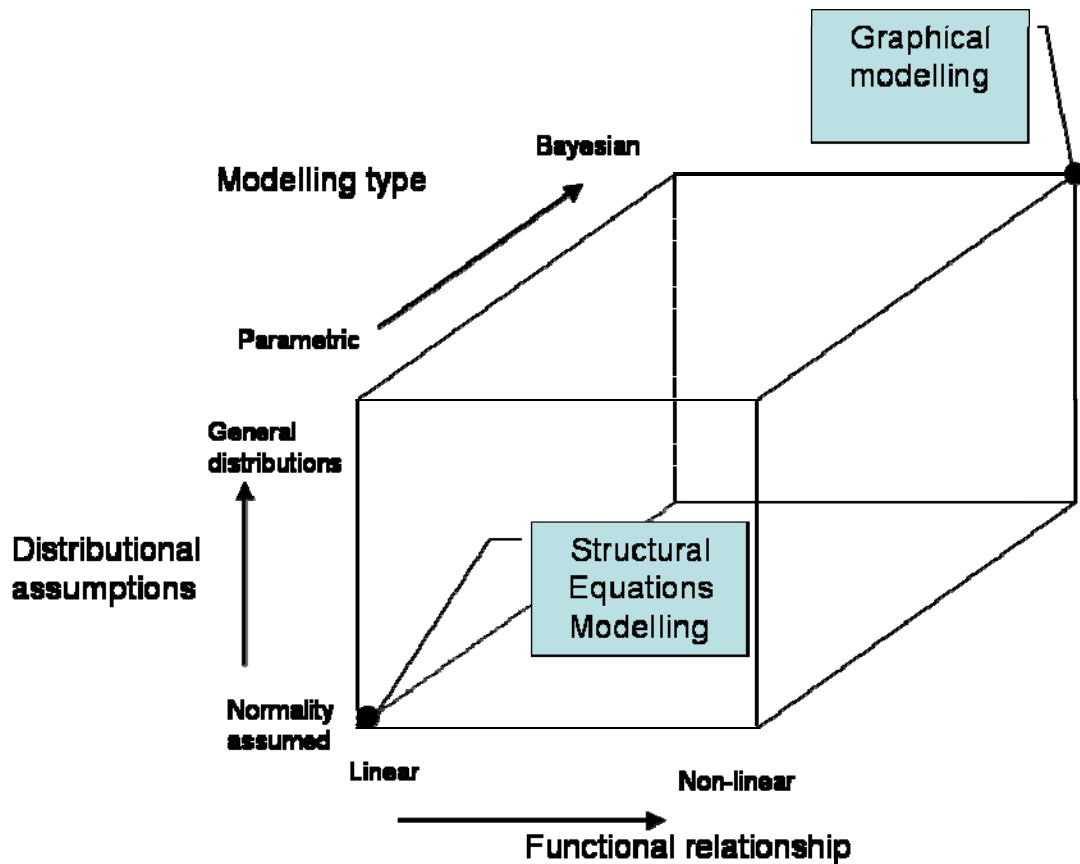


Figure 4: Graphical modelling as an extension of Structural Equations Modelling

On basis of Figure 4 it might seem as if the OPUS project should aim for the use of non-parametric models with general probability distributions and non-linear relationships assumed throughout.

This is not the case for the following reasons:

- such a move would in fact require an overhaul and/or the replacement of most models currently used transportation planning, for which this project neither has the brief nor the resources
- in some cases, the state-of-art model components using linear models are a sufficiently good approximation of reality in the sense that increasing *total* model accuracy would require one to improve *other* model components first
- the possibilities of parametric models have not been exhausted in transportation planning, and parametric models offer many practical advantages, such as economy of the descriptive models and interpretability of the model coefficients

Where the *theoretical* framework will not restrict itself by a-priori assumptions regarding linearity, distribution type, or parameterisation, all practical demonstrations within the project

will have to carefully weigh theoretical advantages against practical requirements such as re-use of existing models/code-bases

#### 4.2.2.1 Graphical models and conditional independence

Consider the graph in Figure 5. Shown are 3 variables: x,y, and z. The graph represents the relationships between the variables x,y, and z.



Figure 5: Independence graph

The *conditional independence structure* implied by the graph is:

$$x \perp y \mid z \tag{38}$$

In words: “x is independent of y given z”.

This deceptively powerful property allows one to:

- Determine if a *decomposition* of a complicated model into simpler submodels exists, and find it if it does
- Encode any structures present in a GAPM in a way that is both understandable and recognizable to a domain expert yet sufficiently rigorous for the needs of statisticians and software developers.
- Separate out structural relationships from measurement relationships

The increase in generality that results from the use of graphical models comes at a price: graphical models are much more abstract than SEMs and can be much harder to operationalise.

#### 4.2.2.2 Separating measurement relationships from structural relationships

In order to capture the relationship between observations and structural variables, the graphical model of the problem area will be enriched with small additional graphs to represent what in the terminology of structural equations modelling are known as *observation equations*. The usual convention is that observable variables are denoted in the graph as rectangles, and latent variables as ellipses. We can add observation relationships to the graph in Figure 5 as shown in Figure 6.

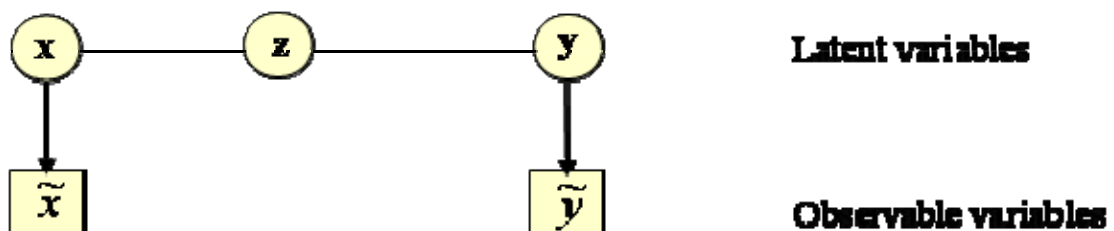


Figure 6: Separation of structural relationships and observations

In this figure, we have separated the nodes of the graph into two levels” latent and observable. In this way one may enrich any graph that encodes a GAPM with observation relationships.

#### 4.2.2.3 A succinct characterisation of graphical models

A succinct description of Graphical Models is given in the following quote found in Murphy (2004):

*"Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering -- uncertainty and complexity -- and in particular they are playing an increasingly important role in the design and analysis of machine learning algorithms. Fundamental to the idea of a graphical model is the notion of modularity -- a complex system is built by combining simpler parts. Probability theory provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data. The graph theoretic side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.*

*Many of the classical multivariate probabilistic systems studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics are special cases of the general graphical model formalism -- examples include mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models. The graphical model framework provides a way to view all of these systems as instances of a common underlying formalism. This view has many advantages -- in particular, specialized techniques that have been developed in one field can be transferred between research communities and exploited more widely. Moreover, the graphical model formalism provides a natural framework for the design of new systems." --- Michael Jordan, 1998.*

In our view, the formal tool of “Graphical Models” presents a unifying structure of a type that is essential in dealing with the problems that the OPUS project addresses. Moreover, operational software exists with which graphical models can be operationalised, as will be shown in section 4.4.

#### 4.2.2.4 Operationalising graphical models

In recent years, a powerful algorithm called the Markov Chain Monte Carlo (MCMC) method, has emerged to numerically construct joint probability distributions *numerically* from a collection of conditional ones, *irrespective* of the analytical form of either the joint probability density function.

As this allows one to cut through the thicket of analytical complexity in order to derive results whilst retaining all features of the distributions, this method has emerged as a workhorse of modern statistics. A description of this method can be found e.g. in Morgan (2000).

#### 4.2.2.5 Pro and contra of graphical models

The theoretical advantages of graphical models is that they

- can handle complicated structure
- are not restricted to linear models
- can distinguish between observable and latent variables in a natural way
- separate structural relationships from observational relationships
- link probability distribution rather than moments

The theoretical disadvantages of graphical models are that they

- may be computationally intensive
- may require non-trivial development effort

Fortunately recent developments show that graphical models can be solved using MCMC methods, and that standard software is available with which to do so.

### 4.3 Identification of probability distributions of observable quantities

Assume that we have a mathematical model of the structural relationships and a link between structural variables and observations, with a certain error structure. The question of how to determine the best estimate of the structural variable that gives rise to an observation is a classical one. The answer depends on the model that describes the observation process; i.e. the process that links the observations to the structural variable thought to correspond to it.

The form of the observational model tends to depend on the discipline of science that one deals with: in the physical sciences (and e.g. in traffic observation) one tends to have relatively straightforward error models. In the life sciences (e.g. Biology, Health) and the social sciences (Economics, Psychology, behavioural transportation modelling) one tends to face substantially more complicated error structures. Of practical interest for the OPUS programme are situations with complicated error-generating mechanisms, as these occur in practice and often cause considerable problems.

An advantage of using graphical models is that one can construct a subgraph that captures the observation process, or even one that links observable instrumental variables to the latent structural variables. Thereby we may separate out the study of how to deal with partially observed variables.

Basic findings on the complications that arise in the identification of variables in the social sciences can be found in Manski (1995), and a systematic treatment of imperfectly observable variables is given in Manski (2003).

#### 4.3.1 Partial identification of probability distributions

We will adopt the framework proposed in Manski (2003). In an ideal world one may construct an empirical distribution  $\tilde{P}(y)$  of an observable variable  $y$ , which is then guaranteed to converge strongly to the probability distribution  $P(y)$ . In this case  $y$  is called *point-identifiable*, meaning that its probability distribution can be determined as a single point in the set of all possible probability distributions of  $y$ .

In the real world, a number of circumstances may conspire to mar this happy state of affairs.

The example given in Manski (2003) to illustrate the possibility of partial identifiability is as follows. Let  $y$  be an observable stochastic variable, and let  $z$  be a binary stochastic variable that determines if  $y$  is observable or not. The probability distribution of  $y$  can then be split as follows:

$$P(y) = P(y | z = 1)P(z = 1) + P(y | z = 0)P(z = 0) \quad (39)$$

As noted in Manski (2003), the observational evidence reveals only the distribution of observable outcomes  $P(y | z = 1)$ , and the distribution of observability  $P(z)$ , and is uninformative about the distribution of the missing outcomes  $P(y | z = 0)$ .

Therefore the most that can be concluded about the distribution  $P(y)$  is that it lies in a *set of probability distributions*  $H[P(y)]$ , called the *identification region*.

$$H[P(y)] = \{P(y | z = 1)P(z = 1) + P(z = 0)\gamma \mid \gamma \in \Gamma_y\} \quad (40)$$

with  $\Gamma_y$  the set of *all* probability distributions for  $y$ .

Note that in the special case that  $P(z = 0) = 0$ ,  $H[P(y)]$  collapses to a single probability distribution:  $P(y | z = 1)$ , and  $y$  is *point-identifiable*.

As  $P(z = 0)$  increases, the identification region increases with it.

#### 4.3.1.1 Relevance of the work on partial identification

The point of view taken in Manski (2003) is interesting and relevant for the OPUS project for several reasons:

1. It provides a very general, yet clear and consistent *logical framework* for missing-observation problems, instrumental variable problems, ecological regressions, response-based sampling, contaminated outcomes, and general treatment of missing data. This framework allows one to instantly categorise earlier work on missing value analysis in terms of the hypothesis adopted, and the contribution to the shrinking of the identifiability region  $H[P(y)]$ .
2. It shows how the adoption of increasingly strong *hypotheses* about the distribution of the variable  $z$  may eventually enable one to shrink, and sometimes (but not always !) collapse the identification region  $H[P(y)]$  to a single probability distribution.
3. It clearly shows that some of these hypotheses are *non-refutable*, because the observations contain no information whatsoever that would allow one to either verify or falsify the hypothesis under consideration.
4. It shows the existence of a large number of cases where the identifiability region  $H[P(y)]$  consists of set of probability distributions; this case is known as *partial identifiability*.
5. It allows one to distinguish clearly between uncertainty introduced by the finite number of observations available, and the intrinsic uncertainty that would persist even with arbitrarily large numbers of observations.
6. It allows one (in certain cases) to obtain *sharp upper and lower bounds* on the amount of uncertainty injected by the partial identifiability into characteristics (such as the mean and the mode) of the probability distribution  $P(y)$ , *without having to adopt any hypotheses at all*.

Of special interest to OPUS is the application of this framework to the case where multiple data-sources are used to reveal the probability distribution of a single quantity, insofar as it might help resolve issues where multiple data-sources seem to contradict each other. We note in this respect that in classical data-analysis, non-refutable distributional hypotheses are made routinely on basis of intuition, engineering judgement, or for practical reasons. We conjecture that a thorough investigation of the underlying (non-refutable) hypotheses of mde in the collection, processing, and interpretation of datasets may shed some light on cases where datasets may seem mutually contradictory.

#### 4.3.2 Analysis of incomplete data under the MAR hypothesis

In real-world data, incompleteness of the raw data (in the sense of individual items of the data matrix being missing), is the rule rather than the exception.

In principle, one could delete every single data record with 1 or more missing items from the dataset. This however has serious drawbacks:

- a) it becomes so much more expensive to collect sufficient data items on which to calibrate models as to be impractical
- b) bias is likely because missing data often means “nonresponse”, or “unobservable item”, which in general is anything but evenly distributed across the population of interest.

A special case occurs when one is prepared to accept the *hypothesis* (note that this *is* a hypothesis, and moreover one that is *not empirically verifiable or refutable!*) that data are Missing At Random (MAR).

If one accepts the MAR hypothesis, then one can use the observed data to “fill in” the missing values. An in-depth treatment of this procedure is presented in Shafer (2000). From a statistical point of view, this method constructs a probability distribution for each data item, which collapses if the item is observed, but does not when the item is missing.

We note that all “pragmatic” methods for imputation of missing data used in the field are special instances of just such a statistical inference method, which unfortunately is not always realised by the data-collecting agency.

## 4.4 Filtering

Filtering is a special branch of statistical estimation usually associated with the processing of time-series observations.

In its most basic form, filtering is about optimally suppressing high-frequency components from a signal containing observations on grounds that those components constitute noise. Filtering may be carried out either in the frequency domain of a signal, or in the time-domain.

### 4.4.1 The Kalman filter

Kalman Filtering is a linear minimum mean-square error (MMSE) filtering process using state-space methods. The two main features of Kalman formulation and problem solution are:

- vector modelling of the dynamic process under consideration
- recursive processing of noisy measurement data.

Assume we have a stochastic process which is defined by its vector of state variables  $\mathbf{X}_k$  at discrete time intervals  $k$ . Assume also that we have a *transition equation* which describes the evolution of the state in time:

$$\mathbf{X}_{k+1} = \mathbf{f}(\mathbf{X}_k) + \varepsilon_{trans} \quad (41)$$

Assume also that the state vector is not directly observable; what we observe is a sequence of variables  $\mathbf{Z}_k$  which are related to the state variable through an *observation equation*:

$$\mathbf{Z}_k = g(\mathbf{X}_k) + \varepsilon_{obs} \quad (42)$$

The problem is to determine the optimal estimates of the unobserved state variables  $\mathbf{X}_k$ .

In the special case where the functions  $\mathbf{f}$  and  $g$  are both *linear*, and the error terms follow a multivariate Normal distribution, the transition equations and the observation equations can be written as:

$$\mathbf{X}_{k+1} = \Phi_k \mathbf{X}_k + \varepsilon_{trans} \quad (43)$$

Assume also that the state vector is not directly observable; what we observe is a sequence of variables  $\mathbf{Z}_k$  which are related to the state variable through an *observation equation*:

$$\mathbf{Z}_k = H_k \mathbf{X}_k + \varepsilon_{obs} \quad (44)$$

the Kalman filter provides the following (optimal) solution:

$$\hat{\mathbf{X}}_k = \hat{\mathbf{X}}_k + K(\mathbf{Z}_k - [\mathbf{Z}_k | \hat{\mathbf{X}}_k])$$

where  $\mathbf{K}_k$  is known as the Kalman gain matrix at time  $k$ , which is defined as:

$$\mathbf{K}_k = \mathbf{P}_{k/k-1} \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_{k/k-1} \mathbf{H}_k^T + \mathbf{R}_k)^{-1}$$

with  $\mathbf{P}_{k/k-1} = \text{cov}(\varepsilon_{trans})$ , and  $\mathbf{R}_k = \text{cov}(\varepsilon_{obs})$ .

It turns out that the Kalman filter can also be applied when  $\varepsilon_{trans}$  and  $\varepsilon_{obs}$  do not follow a Normal distribution, but then the solution may or may not be optimal. In the event that the transition equation and the observation equation are nonlinear, the Kalman filter can still be applied, e.g. by linearising the system equations at every time instant. This is known as the *extended Kalman filter*. A thorough description of the Kalman filter can be found in Grewal and Andrews (2001), or in Gelb (1979).

It is worth noting the form of the updating gain matrix  $\mathbf{K}_k$  in two extreme cases:

1. assume that the observation noise completely dominates any errors in the transition equation. This would be reflected as  $\mathbf{R}_k \gg P_{k|k-1}$ , so that  $\mathbf{K}_k = \mathbf{P}_{k/k-1} \mathbf{H}_k^T \mathbf{R}_k^{-1} (\mathbf{R}_k^{-1} \mathbf{H}_k \mathbf{P}_{k/k-1} \mathbf{H}_k^T + I)^{-1} \approx \mathbf{P}_{k/k-1} \mathbf{H}_k^T \mathbf{R}_k^{-1} \approx 0$ . This would mean that no updates are carried out at all, which is what one would intuitively expect.
2. assume that the observation noise is zero, i.e.  $\mathbf{R}_k = 0$ . This would result in:  $\mathbf{K}_k = \mathbf{P}_{k/k-1} \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_{k/k-1} \mathbf{H}_k^T)^{-1} = \mathbf{P}_{k/k-1} \mathbf{H}_k^T \mathbf{H}_k^{-T} \mathbf{P}_{k|k-1}^{-1} \mathbf{H}_k^{-1} = \mathbf{H}_k^{-1}$ , causing any deviance between  $Z_k$  and  $Z_k | \hat{\mathbf{X}}_k$  to be fed back into the new state estimate without any mitigation. Again this is what one would intuitively expect.

The influence graph for three iterations of the Kalman filter is shown in Figure 7.

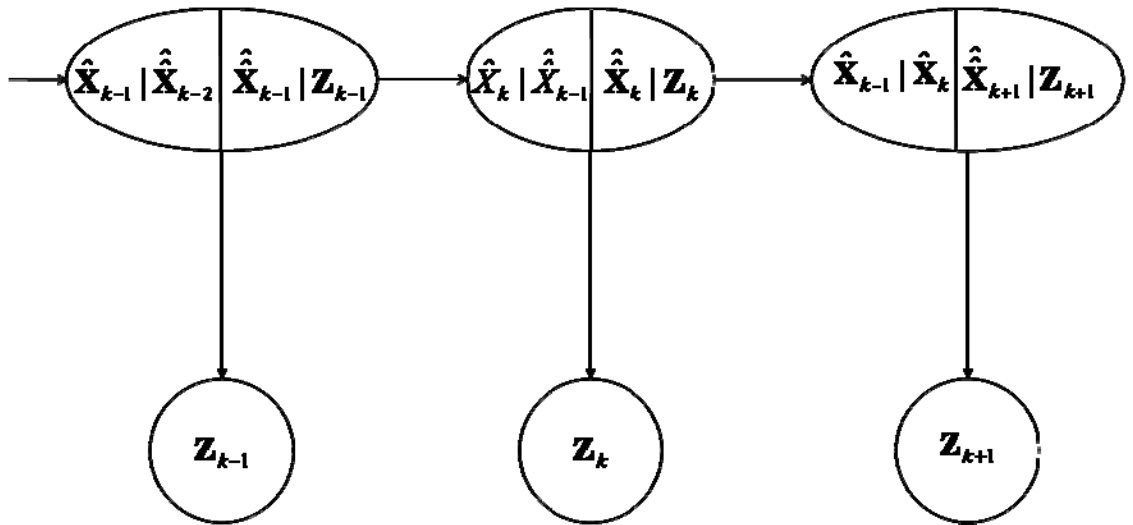


Figure 7: influence graph of the Kalman filter

Shown are the *estimated* vector of state variables at time  $k$ :  $\hat{\mathbf{X}}_k | \hat{\mathbf{X}}_{k-1}$  given the estimated value of  $\hat{\mathbf{X}}_{k-1}$ , the vector of observations  $\mathbf{Z}_k$ , and the *updated* estimate of the state vector  $\hat{\mathbf{X}}_k | \mathbf{Z}_k$  when the observations  $\mathbf{Z}_k$  have been taken into account.

## 4.5 Software identified for use

In this section we will briefly describe the software that we envision for use within work package 2 (theory development), and for communication with project partners that will implement production-quality software.

We note that since the application work tasks deal with data provided by (and sometimes embedded in) external organisations, any attempt to standardise the software used for those demonstrations projects may encounter difficulties.

### 4.5.1 General technical computations and prototyping

In view of its conciseness, expressive power, and widespread use we have identified MATLAB and its Open Source equivalent SCILAB as obvious candidates for general computations and prototyping. Both software packages provide a range of useful computational primitives (such as matrix-vector calculations, basic Linear Algebra, FFT algorithms, signal processing) and for producing graphs. In addition, third-party add-on packages (such as e.g. for Kalman filtering and function optimisation) are available for both.

In view of the conciseness of code in MATLAB or SCILAB, we view working prototypes in these languages (possibly in conjunction with an appropriate specification of the higher-level structure of any data-processing involved by other means) as a suitable vehicle to specify computations in verifiable detail to project partners who will be building production-quality software on basis of the prototypes developed by the research partners.

We note that scientific subroutine libraries exist for most of the high-level primitives provided by MATLAB and SCILAB. Examples would be the well-known Open Source subroutine packages BLAS (Basic Linear Algebra Subroutines) and LAPACK (Linear Algebra PACKage), and the commercially available NAG library.

### 4.5.2 Statistical computations

For specialised statistical calculations we propose to use the packages S, and its Open Source equivalent R. These packages shine when relatively small datasets have to be subjected to complicated calculations.

For cases where large datasets have to be subjected to relatively simple statistical calculations, and for the creation of publication-quality tables, we propose to use SPSS.

### 4.5.3 Graphical modelling

The use of graphical models was prompted on the one hand by its theoretical suitability, but also by the fact that nowadays powerful software exists for the manipulation of graphical models. We mention the BUGS package (see e.g. Spiegelhalter *et al.* (1999), Gilks *et al.* (1996), Spiegelhalter *et al.* (1996), and the BUGS website at <http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>).

This software package has recently been extended with dedicated functionality to deal with problems of spatial statistics, resulting in the GEOBUGS software (see GEOBUGS (2004) and the BUGS website).

## 5. CONCLUSIONS

---

### 5.1 Commonalities noted between successful applications

Combining different data sources is not uncommon in transportation and or health-related applications, but its application seems to be ad-hoc and limited to those situation where an existing probabilistic framework offers convenient support for it.

Examples would be:

- Cases where a known likelihood function is set up which can be adapted to cover more than one data source (as in e.g. Combined SP-RP models)
- Kalman filtering, where various data sources can be accommodated as different parts of an observation vector and merged using their relative variance
- Estimation of mobility for through construction of synthetic populations whose behaviour (at a disaggregate level) comes from one dataset, and whose composition comes from another
- Estimation of O-D matrices on basis of various data sources

The central issue seems to be the construction of a *single likelihood function* that is fed by different but complementary data sources.

Construction of such a likelihood function at first glance seems a formidable task because of the potentially large number of variables and all their interactions (which will be specific to each new situation), and the wide variation in distributional assumptions on the individual datasets. And even if one succeeds in constructing such a likelihood function, it may well prove to be analytically intractable.

The construction of a joint probability density, and therefore a likelihood function may be facilitated, and indeed rendered mechanical through the use of directed acyclical graphical models, as illustrated by the example in which the WINBUGS software was used. In that example the model was specified completely through its graphical model representation and some distributional assumptions. To the WINBUGS software this representation was sufficient to allow it to calculate the probability distribution of the parameters.

### 5.2 Partial solutions identified

On the methodological side, we have:

- identified a formal statistical framework (graphical models), which is likely to be able to deal with the complexity of the problems involved and to be general enough to represent all relevant mathematical and statistical model likely to be of interest
- identified a statistical method that can solve problems formulated in terms of graphical models (the Markov Chan Monte-Carlo method)
- identified a statistical framework (partial identification of probability distributions) which
  - clarifies and generalises the issue of when variables are identifiable from empirical observations in terms of identification regions, and when such identification regions collapse to a single probability distribution (point identifiability) and when they lead to a set of probability distributions (partial identifiability)
  - provides guidance on constructing sharp upper and lower bounds on the expectation and median on partially identified variables
  - clarifies the issue of which additional (unverifiable) hypotheses are needed to collapse the identifiability region to a single distribution

- identified an efficient way of dealing with missing data, provided one is willing to assume the (unverifiable) hypothesis that the data are Missing At Random (MAR)

Regarding the problem domains, we have

- noted that we can represent generic apriori models of a problem domain in terms of graphical models through the use of conditional independence properties that appear to be present in most models in science and engineering
- identified generic sources of error in identification/measurement of models in the life sciences and the social sciences, such as sampling, measurement error, response bias, aggregation, time shifting.

With respect to the operationalisation of the models proposed, we have

- identified an operational software package that can solve problems formulated on graphical models (BUGS)
- identified sources of error in identification or measurement given a local graphical model structure (sampling, measurement error, response bias, aggregation, time shifting)
- identified successful applications of statistical methods to problems that are of the same type as the sub-problems that we identified here (examples from epidemiology, O-D matrix estimation, RP-SP combinations, ...)

In summary we believe that we have presented a viable approach towards the problems that the OPUS project must address, and that we have identified both suitable formal methods and promising software that have proved successful on similar problems in problem domains such as Health and Biology.

### 5.3 Contours of a unified framework

On basis of the examples presented in section 2, the elements of a general approach sketched in section 3, and the tools identified in section 4, we propose the following unified framework/approach for combining multiple datasets in any given problem domain for which a reasonable body of theory exists.

1. Construct a General Apriori Model (GAPM) for the problem domain
2. Translate this into a Graphical Model (GrM)
3. Draw up a joint probability of the relevant variables on basis of the GrM

Next, if the probability functions are *parametric*, then

4. draw up the log-likelihood function for the variables in question as a function of the parameters; this log-likelihood function draws on *all* the datasets
5. and estimate the probability distributions of the parameters from the combined datasets

In case the probability distributions are non-parametric, some further research seems to be indicated.

## 6. REFERENCES

---

- Amemiya, T. (1985) *Advanced econometrics*. Harvard University press.
- Ben-Akiva, M. (1987) 'Methods to combine different data sources and estimate origin-destination matrices' in N.H. Gartner and N.H.M. Wilson (eds) *Transportation and Traffic Flow Theory*, Elsevier, North Holland.
- Ben-Akiva, M., Lerman, S.R. (1984) *Discrete choice analysis: theory and application to travel demand*. MIT press.
- Ben-Akiva, M., Morikawa, T. (1989) Data fusion methods and their applications to origin-destination trip tables. *Transport Policy, Management and Technology Towards 2001* (Selected Proceedings of the 5th World Conference on Transport Research), Yokohama.
- Ben-Akiva, M., Morikawa, T. (1990) Estimation of travel demand models from multiple data sources. *Transportation and Traffic Theory*, (Koshi, M. ed.) 461-476. Elsevier, New York.
- BUGS <http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>
- Bethlehem, J.G., Kertsen, H.M.P. (1986) *Werken met non-respons*. (in Dutch) Ph.D. thesis; University of Amsterdam.
- Bethlehem, J.G., Keller, W.J. (1987) Linear weighting of sample survey data. *Journal of official statistics*. Vol. 3, No.2, 1987. 141-153.
- Bollen, K. (1989) *Structural equations with latent variables*. New York: Wiley.
- Bradley, M.A. (1992) Estimation of logit choice models using mixed stated preference and revealed preference information. *Proceedings of the 6<sup>th</sup> International conference on Travel Behaviour*, Quebec.
- Bradley, M.A., Kroes, E.P. (1990) Simultaneous analysis of stated preference and revealed preference information. Paper for the PTRC 18<sup>th</sup> summer annual meeting. Seminar on Transportation planning methods. September 1990, University of Sussex, England.
- Cascetta, E. (2001) *Transportation systems engineering: theory and methods*. Kluwer Academic publishers.
- Duda, R.O., Hart, P.E., Stork, D.G. (2001) *Pattern classification*. Second edition. Wiley interscience.
- James Fox, J., Daly, A., 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>
- Gelb, A. 1979. *Applied Optimal Estimation*. The MIT Press.
- GEBUGS (2004) <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml>. Website visited 10 March 2004.
- Gilks, W. R, Richardson, S., Spiegelhalter, D.J. (Eds.) (1996) *Markov Chain Monte Carlo Methods in Practice*
- Grewal, M.S., Andrews, A.P. (2001) *Kalman filtering. Theory and practice using MATLAB*. Second edition. Wiley Interscience.
- Kitamura, R., Bovy, P.H.L. (1987) Analysis of attrition biases and trip reporting errors for panel data. *Transpn. Res.* 21A, 4/5, 287-302.
- Lauritzen, S. L. (1996) *Graphical models*. Oxford : Clarendon, 1996
- Logie, M. (2003) The OPUS vision. Internal note. <http://www.opus-project.org/>

- Logie, M. (2001) Specification of data synthesis methodology. Report prepared for Transport for London.
- Louviere, J.L., Hensher, D.A., Swait, J.D. (2000) Stated choice methods. Analysis and application. Cambridge University Press.
- Manski, Ch.F. (1995) *Identification problems in the Social Sciences*. Cambridge, MA: Harvard University Press.
- Manski, Ch. F. (2003) *Partial identification of probability distributions*. Springer series in Statistics, Springer.
- Morgan, B.J.T. (2000) *Applied stochastic modelling*. Arnold.
- Murphy, K. (2004) A Brief Introduction to Graphical Models and Bayesian Networks. MIT webserver: <http://www.ai.mit.edu/~murphyk/Bayes/bnintro.html#repr>
- Ortuzar, J., Willumsen, L. (2001) *Modelling Transport*. Third edition. Wiley.
- Polak, J. (2000a) On the estimation of travel demand in London by combining data from different sources. Unpublished paper.
- Polak, J. (2000b) On the estimation of individual and household trip rates from the LATS and LRTS household surveys. Unpublished paper.
- Polak, J. W. (2000) Analysis of the LATS 2001 Pilot Household Travel Diary Survey, Report to the DETR, Centre for Transport Studies, Imperial College.
- Richardson, A.J., Ampt, E.S., Meyburg, A.H. (1995) *Survey methods for transport planning*. Eucalyptus press.
- Richardson, S. (1996) Measurement error. In Gilks, W. R, Richardson, S., Spiegelhalter, D.J. (Eds.) (1996) *Markov Chain Monte Carlo Methods in Practice*. 401-417
- Richardson, S., Gilks, W.S. (1993) Conditional independence models epidemiological studies with covariate measurement error. *Statist. Med.*, **12**, 1703-1722.
- Sanchez-Gomez, R. (1996) *Analysis of Interpersonal and Intrapersonal Variation in Travel Behaviour in the 1985/86 National Travel Survey*. MSc Dissertation, Centre for Transport Studies, Imperial College.
- Shafer, J.L. (2000) *Analysis of incomplete multivariate data*. Monographs on Statistics and applied probability, no. 72. Chapman & Hall.
- Spiegelhalter, D. J., Thomas, A. Best, N.G. (1999). *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit
- Spiegelhalter, D.J., Thomas, A., Best, N.G. (1996). Computation on Bayesian graphical models. *Bayesian Statistics* 5 pp. 407--425
- Train, K.E. (2002) *Discrete choice methods with simulation*. Cambridge University Press.
- Venables, W.N., Ripley, B.D. (1999) *Modern applied statistics with S-plus*. Third edition. Springer.
- Whittaker, J. (1998) *Graphical models in applied multivariate statistics*. Wiley.
- Zhao, L., Ochieng, W.L., Quddus, M., Noland, R. (2003) Extended Kalman Filter and Map Matching Algorithm for an Integrated GPS/Dead Reckoning System for Transport Telematics Applications. *Journal of Navigation*, 56 (2), 257-275