

OPUS

Optimising the use of Partial information in Urban and regional Systems

Project IST-2001-32471

ITS Programme

Title : **Review and update of the modelling framework**

Author(s) : Axhausen, K. (IVT, ETH Zürich)

Deliverable No. : D5.1
Version : 1.2

Date : Initial version: August 2005

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

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

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

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

TABLE OF CONTENTS

TECHNICAL ABSTRACT	1
EXECUTIVE SUMMARY	2
1. PURPOSE	3
2. BRIEF REVIEW OF THE APPROACH AND DELIVERABLES	4
3. CASE AND FEASIBILITY STUDIES.....	7
4. INITIAL IDEAS FOR EVALUATION	9
5. SUMMARY AND ASSESSMENT	10
6. LITERATURE	11

TECHNICAL ABSTRACT

This deliverable D5.1 is the part of the on-going OPUS internal review process. It takes stock and reviews the progress made so far and outlines the case and feasibility studies planned and started.

EXECUTIVE SUMMARY

This document is Deliverable D5.1 of the Fifth-Framework project OPUS. This review indicates that the OPUS project is on course to achieve its objectives, even if it is slightly delayed in the transformation of the conceptual work into operational software and cases. Still, the recent advances for the case studies should allow the project to catch up with the timetable.

The twin conceptual approach of a generally Bayesian approach supported by appropriate metadata structures will substantially advance the state-of-the-art of modelling in the transport domain, especially through the further development of domain specific samplers. The case studies should provide plenty of material to demonstrate the overall usefulness of the approach.

1. PURPOSE

The OPUS project is developing a general statistical framework to improve the combination of complex spatial and temporal data from survey and non-survey sources. This approach is Bayesian in principle and acknowledges the structural relationship between the variables of interest while accounting sampling and non-sampling errors. The framework will be applied in a series of case and feasibility studies of increasing complexity drawn from the transport and health fields.

The technical annex describes the specific objective of this work package and therefore deliverable as:

- Review of definitions of data flow to check consistency of the items and their interfaces
- Cross check of model assumptions and data situation using the later case study cities as references
- Review of model assumptions with regards to realism and acceptability for the case study cities

based on input from all project partners.

The changes in the timetable and the results obtained since the start of the project have led to a readjustment of these objectives, so as to maximise the contribution of the work to the further progress of the project. The deliverable will focus now on:

- Review of model assumptions with regards to realism and acceptability for the case study cities
- Identification of interactions between the case and feasibility studies
- Outlook towards the evaluation process to ensure, that this can be kept in mind during the conduct of the case studies

The next section will briefly review the available deliverables followed by a discussion of the approach and possible enhancements. The case studies, as envisaged at this time, will be sketched and their potential synergies identified. The final section will discuss the options for the evaluation. A final assessment will round up the deliverable.

2. BRIEF REVIEW OF THE APPROACH AND DELIVERABLES

The following substantive deliverables have been delivered at this time:

- D 2.1 Review of the relevant literature
- D 2.2 Identification of methodology and tools
- D 3.1 Proposals for metadata for generic support of statistical modelling in statistical data bases
- D 3.2 Specification of the extensions of the LATS data base system for the transport domain
- D 4.1 Specification of pilot transport implementation model – Inception report
- D 6.1 Optimising the use of partial information in urban and regional systems

The motivation for the project does not need to be restated at length here, but a brief reminder is appropriate. OPUS addresses one of the central practical difficulties in applied mathematical modelling: the diversity of available data sources and their unknown relative merits and accuracies. The multitude of sources, changing definitions over time, ad-hoc or proper sampling processes produce data streams, which need to be carefully weighted and considered when used. Consider for example the following typical subset for a strategic transport model:

- Land use data from (aged) administrative records
- Transport network data from a Navigation System provider
- Capacities and (partial) flow-speed functions from guidelines
- Travel diary data on travel behaviour from a sample survey of residents, often without a detailed analysis of the non-response behaviour in the population
- Brief screen-line surveys of visitors
- Samples of counts on selected roads and public transport services with unknown sample and non-sample errors

OPUS is constructing an approach, which aims to document and integrate these data streams and the resulting models while integrating structural knowledge about the agent and system behaviour. The twin objectives of modelling, while fully documenting process, data, models and results is reflected in the two streams of work undertaken so far.

Deliverable 2.1 carefully reviews the literature on data fusion, with a particular emphasis on transport applications. It identifies *graphical modelling*, Kalman filtering and explicit modelling of relative size of errors as building blocks. It clearly identifies that the graphical modelling, a Bayesian approach, needs to be guided by general a-prior model (GAPM) to impose structure and the benefit from the accumulated domain knowledge.

These ideas are developed and systematised in Deliverable D 2.2, where they are supplemented by the tools identified to translate this OPUS approach into application. The combination of GAPM and graphical modelling is at the core, while a sophisticated understanding of the data problem informs the discussion. As alluded to above, natural variation, measurement bias, indirect measurement and non-response effects require a complex specification of the graphical models. This discussion highlights the need to make a-priori assumptions about the error processes involved. To guide further thinking a seven level model is applied:

Level		
1	Physical	What is the domain of discourse?
2	Conceptual	What variables do we distinguish? Which do we want to know?
3	Structure	How do they influence each other? Which are observable?
4	Model	How do we specify the relationships?
5	Statistical model	Which measurement errors do we consider, and where?
6	Estimation	Which algorithm can calculate the variables and parameters?
7	Application	

Given the research context and the potential later use of the OPUS approach in the public section, the relevant software tools identified are open source: R, BUGS and SCILAB, while acknowledging that commercial software is required for specialised tasks, such as the VISUM package of the project partner PTV for network algorithms.

Deliverables D 4.1 and 4.2 translate the generic approach into a specific transport application, in which the project aims to estimate an origin-destination matrix for a part of London. The domain specific requirements lead to the integration of the network algorithms of VISUM into a specialised MCMC sampler proposed by Tebaldi and West, 1998. This application has since been finalised (?) and its software will be expanded to much larger problems in the later case studies.

The second strand of the deliverables also covers general as well as the specific issues. Deliverables D 3.1 and D 6.1 develop a new comprehensive approach for the structure of documenting models, data and results in a Bayesian – based modelling environment, as sketched above for OPUS. Special care is given to link the concepts to the prototype implementation using the Unified Modelling Language (UML).

The reasons for the shift from a unified data base to a systematic archiving of the London case study using the NESSTAR/ddi combination are well argued in Deliverable D 3.2.

The same combination is also the basis for the Zürich case study (Chalasan, Schönfelder and Axhausen, 2002). While the flat file structure of NESSTAR does not allow as many sophisticated data base operations as a relational or object-oriented data base would, these can be implemented if needed. The openness, clarity and easy access to the data, also for the general public, is on balance preferable. In addition, this structure allows the project to prototype its model and results documentation approach mentioned above in parallel with NESSTAR, which paves the way for its inclusion into the later version of the ddi Standard, on which NESSTAR is based (Axhausen and Wigan, 2003).

With the results obtained so far, the project has systematically addressed its objective. The combination of a new approach to documentation and the systematic application of Bayesian style approaches to data integration and estimation is very powerful. The main weakness at this point is the time gap between the conceptual development and the implementations and case studies. Still, this gap is in the process of being closed through the on-going work.

At the May 2005 workshop of the work package, two issues were raised with regards to the theoretical and conceptual development: integration of the uncertainty attached to the parameters of the structural models and the description of the history of the data files used.

In the London test application and later in the case studies we rely on travel time estimates from the network models. While the flows predicated from such models are reasonably robust, the time estimates are less so, as the speed-flow relationships embedded in the models are compromises between factual accuracy and numerical convenience. As a result they often receive less attention, than they should. In addition, the network models usually employ only a small number of such functions for ease of use and lack of data, which means the systematic bias has a random component arising from the very diverse local conditions causing differential differences between modelled and observed speeds. The project will have to implement its conceptual framework in a way which properly accounts for such errors.

The metadata concept currently proposed is focusing on the models and their results in a very careful way. The next iteration of the concept will have to extend this care to the description of the history of the estimation data files, as generally a large number of manipulations of the raw data occur before they are used for estimation (imputations, removal of outliers, subsetting, recodings, transformations and rescalings etc.). While one could, in principle, apply the current metadata concept to each and every of these steps, this would needlessly multiply the data sets, and therefore a consistent variable, item and case history mechanism will need to be integrated in the future.

3. CASE AND FEASIBILITY STUDIES

Project schedules and intentions have to be adjusted continuously to reflect the opportunities and constraints arising inside and outside the project; see for example the use of the advanced NESSTAR system for the London case study. The partners will contribute the following cases to demonstrate the OPUS approach with realistically scaled problems.

The Department of Epidemiology and Public Health at Imperial College will merge transport data, emissions measures and health risk data to improve estimates of the impacts of vehicle emissions on public health in the Northampton. The key advance will be the generalisation of the limited measurements to the whole city and the integration of time-of-day and time-of-year dependent effects.

The Centre for Transport Studies (IC) will expand the work started in Deliverable 4.2 to substantially larger areas of London, incorporating further data sources. A second case study is still in the design stage (?), as the partner Transport for London is still considering their options.

Systematica will also be addressing an origin-destination problem in the context of the transport plan for the province of Lombardy. Here substantial origin-destination survey information needs to be integrated with transport models and traffic counts to obtain detailed flow estimates for a potential new river crossing.

The FUNDP will engage in a comparative study of a traditional parametric approach and the Bayesian approach of the project. The problem at hand is the generation of a large scale artificial sample of agents for later micro-simulation work. Using census information FUNDP will create these samples and carefully compare the results.

The World Health Organization will define potential areas of application for OPUS methods within the health domain. An area of relevance to the development and utilization of OPUS, that is particularly important to the health sector, is the modelling of spatio-temporal exposure to health data.

ETH will undertake two case studies. The first case study addresses the combination of various data sources to obtain a consistent and richer estimate of travel. Currently three sources provide information about the number of leisure excursions and day trip (national travel survey (Mikrozensus Verkehr 2000), national income and expenditure survey (Einkommens und Verbrauchserhebung 2000) and the privately funded survey Schweizer Reisemarkt (Swiss Travel Market). These three are inconsistent in their scopes, partially in the object definitions and their sampling processes. The aim of this case study is to integrate the partial information available in each to obtain a joint estimate of the distribution of trip making for excursions.

The second case study is based the strategic cantonal traffic model (about 900 zones), which has been developed at the IVT for the year 2000. Using both count information, but more importantly speeds from a floating car study the 2000 matrix will be updated. The challenge is the integration of the variance of the counts and the model error in the speed data, as the floating car data is generalised to the whole network using a suitable spatially-aware regression model.

The possible third case study is similar in spirit to the work at FUNDP. If possible, the IVT plans to extend an existing artificial sample with information about mobility tool ownership (car, bicycles, public transport season tickets).

In spite of the range of the case and feasibility studies, one common theme is identifiable: the interaction between network models and various behavioural data. The wish to apply the OPUS approach to large scale networks will require the scaling and further development of the Tebaldi and West sampler to realistic network sizes. The expertise of the partner PTV will be put to best use here.

Otherwise, the range of topics will help to demonstrate the broad scope of the approach, while it focus on transport question will help to advance the application of Bayesian methods in this field of study.

4. INITIAL IDEAS FOR EVALUATION

The on-going evaluation of any project work is essential for the success of any project. The focus so far in the OPUS was on the conceptual development, but with the shift to the case studies other issues are moving to the foreground. In advance of Deliverable D 12.1 Evaluation Plan this section will discuss the shape of the evaluation process. As discussed in the Zürich Consortium Meeting in April the evaluation will need to address a range of questions:

- A qualitative assessment how far the initial objectives of the project have been achieved
- A qualitative and quantitative assessment of the contribution of the OPUS approach in the case studies

The first part of the evaluation raises no specific challenges, but the second one does due to the objectives of the OPUS project. The motivation behind OPUS is the belief that the systematic combination of data sources improves modelling by providing both more precise estimates, but also by providing estimates of the variances involved. This second element is not normally provided in transport modelling and the project cannot fall back on established approaches. The case study evaluations should therefore have three elements:

- Match against independent global indicators and distributions, as suitable for the case (trip length distributions, observed flows, volume of journeys etc.)
- Differential improvement of the modelling results as a function of the amount and range of data and data sources added to the original data sources.
- Qualitative assessment of the submodels and of the estimates of the variances of the parameters and variable values.

The second element will require the case studies to systematically vary the additional information brought to bear in the case study, so that their impact on the both the match against the global indicators as well as against the variances can be assessed. It will also require the case studies to hold back certain data items, so that they retain independent global indicators. The next iteration of the case study descriptions should indicate these data sets and the strategy for the differential testing which is appropriate for the individual case.

5. SUMMARY AND ASSESSMENT

The review indicates that the OPUS project is on course to achieve its objectives, even if it is slightly delayed in the transformation of the conceptual work into operational software and cases. Still, the recent advances for the case studies should allow the project to catch up with the timetable.

The twin conceptual approach of a generally Bayesian approach supported by appropriate metadata structures will substantially advance the state-of-the-art of modelling in the transport domain, especially through the further development of domain specific samplers. The case studies should provide plenty of material to demonstrate the overall usefulness of the approach.

6. LITERATURE

Axhausen, K. W. and M. R. Wigan (2003) Public use of travel surveys: The metadata perspective, in P. Stopher and P. M. Jones (eds.) *Transport Survey Quality and Innovation*, 605-628, Pergamon, Oxford.

Chalasan, V. S., S. Schönfelder and K. W. Axhausen (2002) Archiving travel data: The Mobidrive example, *Arbeitsberichte Verkehrs- und Raumplanung*, 129, Institut für Verkehrsplanung, Transporttechnik, Stassen- und Eisenbahnbau (IVT), ETH, Zürich.

Tebaldi, C. And M. West (1998) Bayesian inference on network traffic using link count data, *Journal of the American Statistical Association*, **93** (442) 557-576.