OPUS

Optimising the use of Partial information in Urban and regional Systems

Project IST-2001-32471

IST Programme

Title :	**Zurich Case study 1 : Combining travel information from multiple data sources**

Author(s) :	Chalasani, V.S.	(ETH Zurich)
	Axhausen, K.W.	(ETH Zurich)

Deliverable No. :	D9.2
Version :	1.1
Date:	Initial version: April 2006 Revised:
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 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

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

TABLE OF CONTENTS

TECHNICAL ABSTRACT

This deliverable D9.1 is a result of Work package WP9 of the OPUS project. The aim of this Deliverable is to apply the OPUS methodology that has been developed in the earlier stages. Travel information collected in a travel survey is not complete in nature and its extent of applicability seldom limited to the environments with similar prevailing conditions. Additionally, generalisation of the travel patterns from a single travel survey has not proven as an efficient due to the lack of external evidence. From the vast Swiss travel survey catalogue, three independently conducted travel surveys selected to combine them are:

- Travel module of “Household Income and Consumption survey 1998” (EVE 1998)
- Microcensus 2000 (MZ 2000)
- Travel market Switzerland 2001 (SRM2001).

The central issue in combining the travel information from different sources is to identify the common thread that can be used to combine them to draw a generalised inference. The common travel information in the form of “Excursions”, a home-based round trip that lasts a minimum of 4 hours and a maximum of 24 hours with distance to the destination is at least 25 km, has been derived from the preliminary data mining.

Probabilistic reasoning through Bayesian inference was considered as the platform to combine the excursions information observed in the selected travel surveys. Assuming the data sources as independent, a set of four statistical models in the form of linear regression models was constructed. To account the unobserved effects outside the model, random coefficients in the form of a two level random-effect parameters have been included to the model. An exclusive model was developed based on the assumption that the selected surveys were conducted simultaneously and hence the excursions information was treated as observed information from a single multi-mode multi-period survey. The designed statistical models were constructed as graphical models and tested using the MCMC simulations with Gibbs sampling technique in WinBUGS environment. Based on the posterior distributions of the model parameters, new excursions distributions were calculated for both individual data sources and generalised case. These derived excursions distributions were found significantly different to that of the reported excursion distributions. The generalised excursions distributions based on the joint posterior distributions is the main outcome of this study.

EXECUTIVE SUMMARY

This document is Deliverable D9.1 of the Fifth-Framework project OPUS. The OPUS project aims to develop and demonstrate statistically sound methods of combining datasets that each provides partial or incomplete information on a single complex underlying quantity of interest or set of such quantities.

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

Objectives of Work Package WP9

The objectives of this work package are to:

- Prove the methodology in a practical context with real data
- Capture the benefits for the OPUS project objectives, from the information and modelling infrastructure being developed in London
- Provide concrete improvements to the policy monitoring and modelling methodology available in Zurich
- Demonstrate, conjointly with the pilot application in London, the transferability of the OPUS methodology across different national and application contexts.

Objectives of Deliverable D9.2

The objectives of this deliverable are to:

- Review the theoretical framework of the OPUS methodology developed so far
- Structure the Zurich case study with a set of problems covering all aspects of the OPUS methodology
- Identify the relevant data sources, collect and process the data.
-

1. INTRODUCTION AND FRAMEWORK

1.1 The Zurich case study

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.

Each partner will participate in the technical work of OPUS, contributing to methodology development, commenting on reports from others and participating in User Forum and technical meetings. At the appropriate point in the programme, each will undertake a substantial review of the case studies to apply the methodology in their area. Two of those case studies in the transport sector have been planned for:

- Zurich case study 1: Integrating travel information from multiple sources

- Zurich case study 2 : Update the Zurich cantonal O-D matrix

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.

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 (through prior distributions) and modelling assumptions;
- Additional information is supplied and used to update the model;
- The updated model can be used to provide coherent estimators (with the estimates of reliability) for any area that it covers, including combinations of factors for which no data were actually observed. For example it could provide estimates for passengers leaving a particular railway station in a period when no survey information was collected, but overall passenger loading is known;
- As well as parameter estimates, it is possible to use to model to synthesize simulated data sets that demonstrate behaviour of the system, including its variability;

There is scope within the project for the reliance on Bayesian methods to be supplemented with other techniques without altering the general vision. OPUS has implemented its approach using mainly a MCMC (Markov Chain Monte Carlo) simulation techniques, which are already widely used in statistical studies.

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

1.6.1 Workpackage 9

The relevant objectives of workpackage 9 are:

- Prove the methodology in a particular context with real data.
- Capture benefits for the OPUS project objectives, from the information and modelling infrastructure being developed.
- Provide concrete improvements to the policy monitoring and modelling methodology available
- To demonstrate, conjointly with the pilot application in London, the transferability of the OPUS methodology across different national and application contexts.

1.6.2 Zurich case study1

The Zurich case study 1 (Combining the travel information from multiple sources) focuses on updating the similar information observed from disparate sources. Each set of observed information gets benefited from other independently observed information. The use of external information as an evidence to the

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 undertake a review and synthesis of the relevant domain literature covering successful solutions to problems similar to the ones being address in OPUS. This is the main objective of the current Deliverable. This review contributes to the project both to clarifying the nature of the problems being addressed by OPUS and by identifying generic features of these problems that will be tackled in the theoretical framework. The review also begins the processes of identifying and formulating the elements of the theoretical framework that OPUS will use to deal with these problems.

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

The data sources and Zurich case study 1 methodology will be presented in section 2. Primary analysis and data preparation will be covered in section 3. Section 4 describes the model's framework with variable description to calculate the joint posterior distributions of the. In section 5 results will be described along with necessary interpretations and the new travel demand estimation suing the updated Bayesian inference will also be explained.. In section 6 we will draw the observations together, and outline the conclusions.

2. DATA SOURCES AND METHODOLOGY

2.1 Background and objectives

Zurich case study is designed to test the OPUS methodology developed earlier. As mentioned earlier this work package (WP9) comprises of a set of case studies:

- Combining travel information from multiple sources
- Updating the Zurich cantonal O – D matrices using floating car data

This report covers only the first case study, i.e., Combining travel information from multiple sources.

This study is about the use of an alternative, Bayesian, statistical inference in combining travel information from multiple sources. Travel demand estimation is an essentially dynamic process, in which any individual data collection effort takes place in a context of continuously increasing knowledge. New travel information emerges not only from travel surveys but also from other relevant studies such as volume counts, land-use studies, etc. Furthermore, it is unlikely that a single attempt will provide a definitive conclusion on travel demand, whether the observed demand is erroneous, or it does not represent the actual travel demand. It therefore seems vital to acknowledge the *context* in which a study or an attempt takes place, and to emphasize that the data from a single attempt *add* to available evidence, rather than form the basis for travel demand estimation. The Bayesian method is naturally dynamic, in that prior distributions of belief regarding the trip-making differences may be modified by new data in a formal way by using Bayes theorem.

To combine the travel information observed from the selected travel surveys conducted by different sources, following assumptions have been made in this study:

- Each data collection attempt is *unique* and *random* in nature.
- No single attempt fully explains the travel demand distributions.

2.2 Data sources

Swiss travel survey history is only three decades old, but includes many high quality surveys ranging from simple road-side vehicles counts to the complex travel diaries. In addition to the independent travel surveys, several longitudinal travel surveys in series have been conducted. Three of such longitudinal surveys for which data is:

- Microcensus – Swiss national household travel survey
- Travel module of “Household Income and Consumption survey”
- Travel Market Switzerland

Three travel surveys, one from each series conducted during 1998 and 2001, have been chosen for the first case study. Description

2.2.1 Microcensus (MZ):

Microcensus is the national household travel survey of typical daily travel performed by people in households all over the Switzerland (SFSO, 2001). The travel survey is conducted every five years by Swiss Federal Office for Spatial Development (ARE) and the Swiss Federal Statistical Office (BFS), and collects information about travel by all modes of transportation. It is the only authoritative source of national data on the amount and nature of daily personal travel, and the only source that allows to assess how travel has changed for the nation as a whole. The 2000 survey is the sixth in a series that began in 1974, and was conducted in 1979, 1984, 1989, and 1994. The 2000 Microcensus incorporated state-of-the-practice survey methods, including pre-notification letters, telephone recruiting, and computer aided telephone interviewing (CATI) to aid in the collection of the data. The 2000 Microcensus was conducted seven days a week, including all holidays, from January 2000 to December 2000. Survey data were collected from a sample of 29407 persons from 27917 households making over 100,000 daily trips. Both urban and rural areas were included. All trips made during a pre-assigned 24-hour travel day by each target person six years of age and over in the sampled households were included in the survey. Details about the purpose of every trip, means of transportation, trip time and duration for all the day trips are observed in the survey. Apart from the daily mobility, long duration travel (journeys with over night stays) and long distance travel (air trips) are also observed in the 2000 Microcensus.

2.2.2 Travel module of "Household Income and Expenditure survey" (EVE)

Travel module of EVE 1998 was conducted by the Swiss Federal Statistical Office (BFS) for information on the behaviour of Swiss people with regard to travel. The previous survey, conducted in 1987/88 again as a travel module of EVE. Travel module of 1998 EVE is not a daily mobility survey. It contains information on both the private and business purpose within day, and long duration travel. Approximately half of the main study (EVE 1998) sample was selected randomly (4430 households and 4430 persons) and Computer Aided Telephone Interview (CATI) technique was applied in collecting the data. Different reference periods were set to reduce the burden on respondents. Data collection process was carried out for one year. In a way, this survey can be considered as a panel survey with different questionnaire. A major advantage of this survey is the information on households and persons that can be inherited from the main survey. This follow-up survey with a different and specific purpose was used as a cross check for the main survey with regards to the basic information. The latest travel survey in this series was conducted in the year 2003 (EVE 2003), also as part of household income and consumption survey (Simma, 2005).

2.2.3 Travel market Switzerland (SRM)

Travel Market Switzerland (SRM) is a survey with regard to the travel behaviour of the Swiss population carried out by Institute for Public Services and Tourism at the University of St. Gallen (IDT-HSG). SRM 2001 is the 15th survey in the series that was started in 1972. A total of 5'004 persons from 1'965 households were questioned about their leisure activities. A paper-pencil based self-administrated questionnaire was used to observe the leisure activities over a period of three months. This panel survey was conducted over a year in four consecutive 3 month reporting periods. SRM 2001 is still an important database with regard to leisure journeys by the Swiss population.

An extensive comparison of these three surveys was conducted as a part of internal research at IVT; ETH Zurich (Chalasani and Axhausen, 2003). Table 1 shows the sample properties of the selected surveys.

Table 1 Sample properties

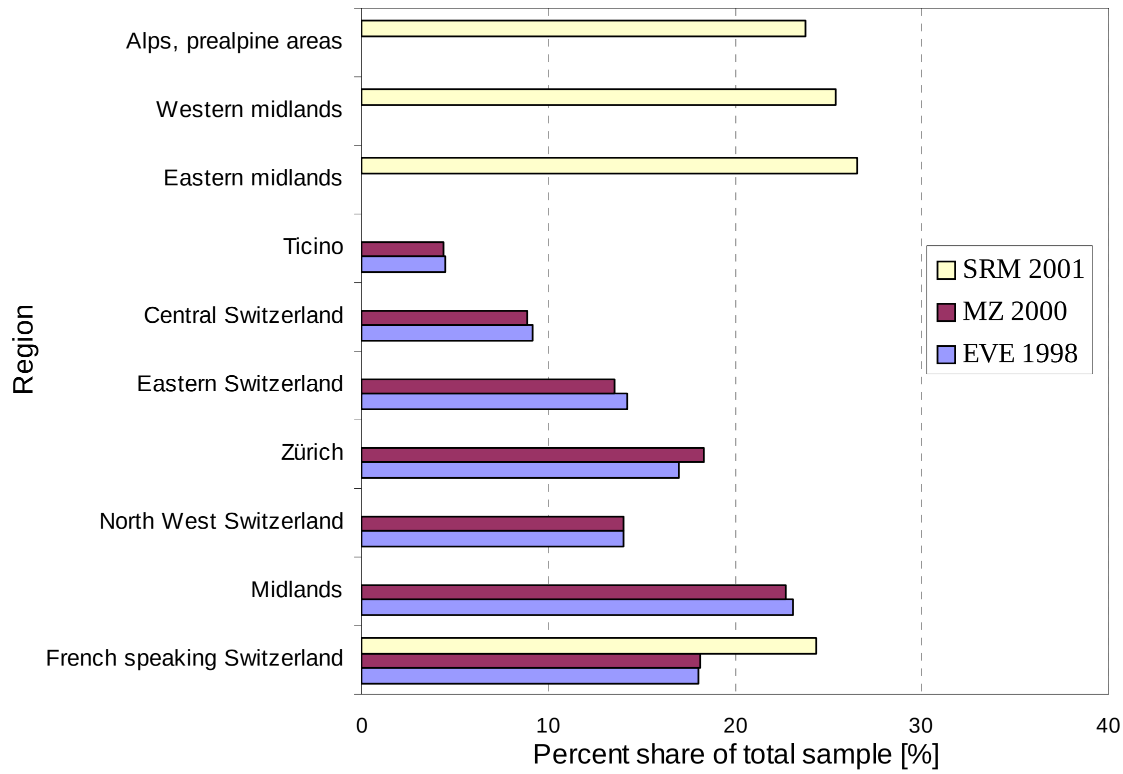
Survey	Target population	Sample size	Expansion factor
EVE 1998	5,875,987[a]	4,460	1’317
MZ 2000	6,815,401[b]	29,492	231
SRM 2001	5,837,000[c]	3,150	1’853

a: Persons above 15 years b: Persons aged 6 and above c: Excluding the Ticino

Source: Chalasani and Axhausen, 2003.

Figure 1 shows the geographical distribution of the samples at NUTS-II level.

Figure 1 Geographical distribution of the samples



Source: Chalasani and Axhausen, 2003

The sample properties shown above make clear that the selected surveys differ both in structure as well as in their geographical distributions. This can be treated as an early sign of the diversity that the selected surveys exhibit.

2.3 Methodology

This section describes a generic methodology with a specific framework and a procedure outlining a set of steps to combine the leisure travel information. There are few many studies on combining information from multiple sources in the transport domain. The history of combining information from multiple sources does not reveal any well defined methodology. Most studies follow a simple and generic methodology, i.e. – select the information in common, identify the relationships, model the information with knowledge as prior, and estimate the modified information using the posteriors. This generic methodology was applied to the current problem, i.e. combining leisure travel information from multiple sources. The following distinct objectives were pursued:

- A common travel demand object observed in the selected surveys
- A set of common travel demand variables observed in the samples by diverse measurement approaches
- Interactions within the travel demand variables and the measured travel demand
- Possible models forms to integrate the different data sources
- Estimating the travel demand based on the posterior distributions as final model output

This study deals the problem of combining leisure travel information in three phases, which can be described as follows:

- Pre-modelling: This initial phase is intended to prepare the data for modelling. Analyses look at the similarities and differences in travel demand and its associated variables that were found in common across the multiple sources.
- Modelling: In this crucial phase, different models will be tested and
- Post-modelling: Output obtained in the form of posterior distributions will be used to estimate travel demand. The interpretation of the results obtained is also an important issue in this last phase.

An efficient methodology in the form of a framework that covers the three phases and the work planned is essential. Figure 2 shows the framework that includes the important stages as well as the software tools used. This study strictly followed this framework to combine the leisure travel information from the selected three surveys. The sequential working steps are:

Step 1: Identify and select the information, both travel demand and its associated, in common across the three selected surveys.

Step 2: Analyse the selected common information for similarities and differences in variable categorization, range and

Step 3: Harmonise the travel associated variables with respect to categories, range and interactions.

Step 4: Prepare the data in “S-Plus” format for input to WinBUGS

Step 5: Identify the interactions among the travel associated variables and their overall interaction with travel demand, the dependent variable.

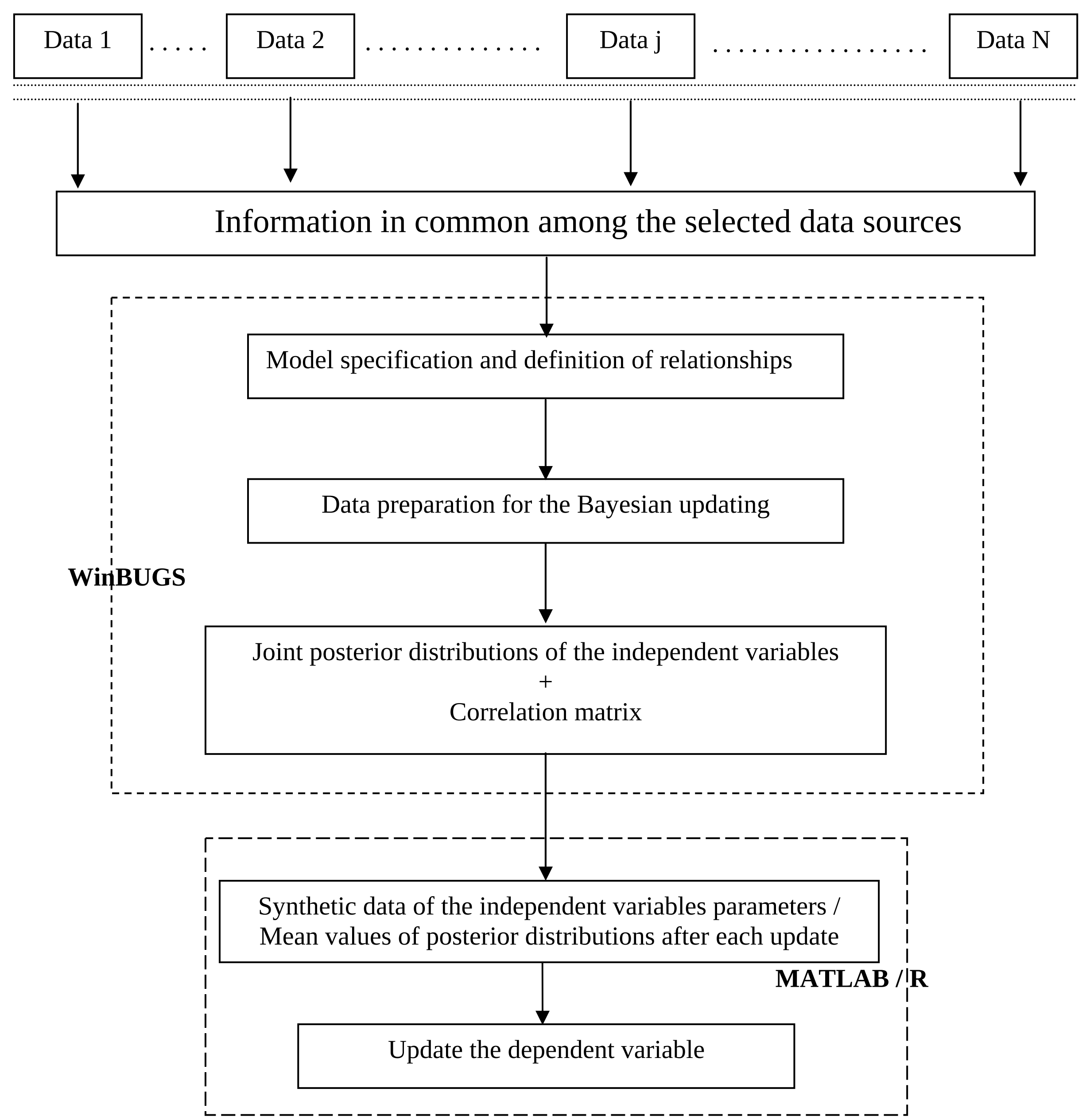
Step 6: Develop a random effect Bayesian liner regression model covering all identified interactions. In the process, choose the parameters for the independent variables, their distributions and priors.

Step 7: Validate the model, load the data, specify the priors and update the WinBUGS model for the maximum number of iterations. Check for convergence with multiple chains.

Step 8: Calculate the posterior distributions of the random effect regression model parameters and the associated correlation matrices.

Step 9: Using the posterior distributions and associated correlation matrices, generate synthetic data by repeated random draws for all model variables.

Figure 2 Framework to integrate leisure travel information from multiple sources



Step 10: Estimate the new travel demand distribution for each survey as well as the joint travel demand distribution. Use the model form implemented in WinBUGS and re-calculate the travel demand using the synthetic data of the parameters and the in

Step 11: Compare the estimated travel demand distributions of the individual travel surveys with that of the joint travel demand distribution to interpret the variations in travel demand in each travel survey with respect to various explanatory variables such as income, car ownership, etc.

2.4 Summary and conclusion

In this chapter we have briefly discussed the data sources considered in this study for the OPUS: combining leisure travel information from multiple travel surveys. The data sources exhibit diversity with respect to the sample properties. A set of distinct objectives identified were specified. The solution was structured into three phases: pre-modelling, modelling and post-modelling phase. A framework satisfying the objectives and covering the work planned in three phases was proposed. A procedure with its steps was described.

3. PRE-MODELLING: PRIMARY ANALYSIS AND DATA PREPARATION

As specified in the proposed framework for integrating information from multiple sources, the first step is to identify the common object shared by the data sources. A preliminary analysis of the three data sources was conducted in two stages. First, to identify the common object, and second, to prepare the data for modelling the information integration. This chapter describes the three surveys in detail and provides information about the common dependent variable.

3.1 Observed information

To start with, EVE 1998 is a module of the regular Household Income and Consumption (EVE) Survey. A sub-set of the sample households from the main EVE survey was used for the travel module. One respondent from each household of age 15 and above was questioned about leisure and business trips with or without overnight stays that were made during the preceding months. Microcensus 2000 (MZ 2000) is a one-day trip diary, in which 29’407 persons of age 6 and above from 25’987 households were asked to report all trips including stages of distance above 100 m made on a pre-defined day. Travel Market Switzerland (SRM 2001) is a panel survey on leisure travel, in which 5’004 persons of age more than 14 from 1’965 households were asked to report all the leisure travel

3.2 Common object

Though the travel information collected in the selected surveys is substantial, it is diverse mainly due to survey characteristics such as context, methodology, data retrieval technique, reporting periods, etc. The following information is in common among the three surveys:

Household information: Household size, household location, monthly household income

Personal information: Age, gender, car availability, public transport season tickets, working status

Travel information: Within-day leisure travel information at different levels of detail such as stage, trip, journey, etc.

Both EVE 1998 and SRM 2001 have used survey specific definitions for within-day leisure travel. In EVE 1998, “*private excursions*” are same-day round trips with no professional or commercial involvement and lasting at least 4 hours away from usual environment. As per SRM 2001, “*Excursions*” are defined as journeys by private persons with no overnight stay and away from everyday life with a distance to destination more than 30km. Covering both definitions, this study has come up with the following definition for “*Excursions*”, the travel demand modelled here:

- Home based round trip
- Leisure travel
- Maximum duration of 24 hours

- Minimum duration greater than 4 hours or distance to the destination more than 25km

Leisure travel observed in both EVE 1998 and SRM 2001 meet the above criteria, thus were directly used in the model. In the Microcensus 2000, leisure travel was observed at trip level. Using the reported origin – destination information, home-based journeys for leisure were identified and then those with distance to destination of above 25 km were considered. Table 2 shows the excursions frequency and its variance among the selected surveys.

Table 2 Excursion frequencies of the three travel surveys

Description	EVE 1998	MZ 2000	SRM 2001
Sample	4’430	29’407	5’004
Reporting period (days)	14	1	365
Excursions as a share [%] of all reported travel	17	02	43
Share of respondents reporting excursions [%]	30	04	64
Average excursions frequency (per person per year)	7.38(0.859*)	9.34 (0.025*)	3.30 (2.025*)

*: Variance

The variations in the average excursions frequencies can be explained by:

- Differences in the reporting periods
- Differences in the travel information aimed for
- Linear extrapolation of the average excursions frequencies

Another interesting issue is the more or less constant ratio of percent of excursions in the reported travel information to the percent of sample reported.

3.3 Primary analysis for data preparation

Data collected in three selected surveys was analysed to understand the similarities and differences in categorisation, range and interactions. It also provides the knowledge on the associated information, which further helps in the initialisation of the model parameters and their priors. Primary analysis and data preparation was performed in the following sequence:

- A simple analysis of the travel related variables to understand the similarities and differences in their categorisation
- Extract common information via non-common variables
- Identify the common categories and harmonise them

- Perform a simple statistical analysis to understand the interactions among the travel variables

- Transform the data into “S-Plus” format, one of the standard input formats of WinBUGS.

A noteworthy point about the common information is, there were differences in scope as well as in the measurement instruments. For instance, income information was observed in each survey. In EVE 1998, detailed information about all sources of income at was reported. In case of MZ 2000 and SRM 2001, a categorical question was framed to observe income. Another difference in income observation was the level, in MZ 2000 and EVE 1998 household income was reported. Whereas in SRM 2001 income was reported at the household head. A part of the common information was observed indirectly. This indirectly observed information was extracted and harmonised to that of directly observed information. For instance, there was no direct question on the car ownership at person level in SRM 2001 and EVE 1998. Car availability information in these surveys was rather extracted from the information about number of cars per household.

Household size: Household size is the total number of permanently residing household members. An open question was posed to collect the household size information. Categories of the household information were used as

Income: Income is an important variable in travel demand modelling because of its compound influence on travel behaviour. As mentioned earlier in this chapter, income was collected using different measuring instruments at different levels. Income at household level (Household income) was reported in EVE 1998 and MZ 2000, whereas income at person level (Household head income) was observed in SRM 2001. Detailed income information was collected in EVE 1998 main surveys and income in categories was provided for both MZ 2000 and SRM 2001 respondents. Table 3 shows the common income categories used in this study.

Table 3 Income categories [CHF]

Category	EVE 1998*	MZ 2000*	SRM 2001+
1	Up to 2’000	Up to 2’000	Up to 2’449
2	2’001 – 4’000	2’001 – 4’000	2’450 – 3’249
3	4’001 – 6’000	4’001 – 6’000	3’250 – 4’049
4	6’001 – 8’000	6’001 – 8’000	4’050 – 4’849
5	8’001 – 10’000	8’001 – 10’000	4’850 – 5’649
6	10’001 – 12’000	10’001 – 12’000	5’650 – 6’449
7	12’001 – 14’000	12’001 – 14’000	6’450 – 7’249
8	Above 14’000	Above 14’000	7,250 – 8’049
9			8’050 – 9’649
10			Above 9’649

*: Monthly household income +: Monthly household head’s income

Land-use type of household location: This is an ideal example of indirect information extracted from outside the travel information reported in the three selected surveys. Swiss Federal Office for Spatial development (ARE, 2002) has categorised Swiss municipalities based on their access to public transport. Table 4 shows the traffic based location categories.

Table 4 Traffic based location types

Category	Description
1	Big city
2	Small city and town with train connections
3	Town without train connections
4	Agglomeration
5	Small town and agrarian municipality
6	Tourist municipality

The age of the respondents was reported in years. Table 5 shows the age categories used in the model. The age ranges were chosen in consistent with the studies on mobility biographies, in which mobility has been analysed at different phases of the human life cycle (Beige and Axhausen, 2005).

Table 5 Age of the person

Category	EVE 1998	MZ 2000	SRM 2001
1	15 - 17	6 – 17	0 – 17
2	18 – 24	18 – 24	18 – 24
3	25 – 44	25 – 44	25 – 44
4	45 – 64	45 – 64	45 – 64
5	65 and above	65 and above	65 and above

Lower limit of category 1 varies due to the differences in the original surveys

Gender of the person is reported in all three surveys, but with different category coding. A common category coding of Female (0) and Male (1) has been adopted here.

Car availability is an important measure of mobility tool ownership. Car availability has direct impact on the mode of transport used for various activities. In EVE 1998 and SRM 2001, car availability information was extracted from two variables: driving license and the number of cars per adult with driving license. Average number of cars per adult with driving license was calculated using directly observed number of cars in the household and number of adults with driving licenses. Car availability information of the three surveys was measured with different precision. This information was harmonised and common categories were drawn. Table 6 shows the categories used for car availability information.

Table 6 Car availability

Category	EVE 1998 and SRM 2001	MZ 2000
1.00	Driving licence = yes and no. of cars per adult>0.75	Car availability = yes
0.50	Driving licence = yes and 0.25> no. of cars per adult<0.75	Car availability = available after arrangement
0.00	Driving licence = No or Driving licence = Yes and no. cars per adult <0.25	Car availability = No

Public transport season tickets are prepaid access cards to the corresponding public transport. It is reasonable to assume that prepaid access to the public transport has a strong influence on travel patterns. Information on the ownership of public transport season tickets was observed uniquely among the surveys. This part of common information observed at different levels and with different perspectives can be converted in a much more meaningful way of motivation to use the public transport. This study treats the mobility tool ownership in the form of a public transport season ticket as a motivation to use the public transport. Switzerland has an integrated yet complex structure of public transport season tickets. With the mode type and area the season tickets vary from a simple city day card to an annual season ticket for whole Switzerland. 100% or fully motivate in case of Swiss national season ticket and 0% or not interested in case of no season tickets. Each type of season ticket was weighted and finally a set of 7 season ticket types were commonly found. Table 7 shows various public transport season tickets and their weights as motivation to use public transport.

Table 7 Public transport season tickets as motivation to use public transport

Category	EVE 1998	MZ 2000	SRM 2001
1.0	Percent of travel made by transit>75	GA = Yes	GA = Yes
0.6	-	Half-fare card + yearly season ticket for the local public transport	-Half-fare card + season ticket for the local public transport
0.5	50<Percent of travel made by transit<75	-	-
0.4	-	Half-fare card + monthly season ticket for the local public transport	Half-fare card + route pass
0.3	-	Half-fare card + weekly season ticket for the local public transport	-
0.1	-	Half-fare card	Half-fare card
0.0	Percent of travel by PT < 25 or not used	None of the above	None of the above

GA: Swiss national season ticket

Working status of the person also influences travel behaviour. Many studies have been carried out on work travel and its influence on overall travel behaviour. Information on working status is normally observed in two ways, first, a simple categorical question on working status with categories like full-time, part-time, etc., second, as a question about the number of working hours per week. The observed working status information in three surveys differs from each other. A simple three category structure was constructed for modelling purposes and the information from each survey was categorised accordingly. Table 8 shows the description of observed working status information categories.

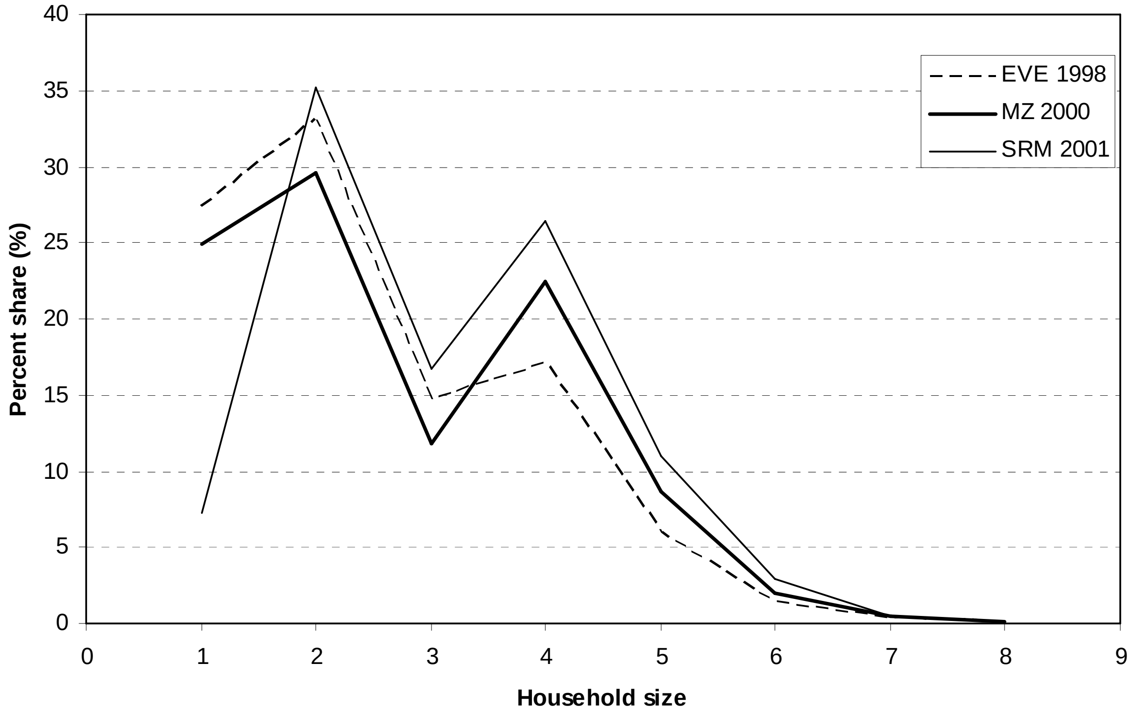
Table 8 Working status

Category	EVE 1998	MZ 2000	SRM 2001
1.00	No. of weekly working hours >30	Working fulltime	Employed
0.50	30<No. of weekly working hours <10	Working part-time	-
0.00	No. of weekly working hours <10 or unemployed	Looking for job or unemployed or retired or house work	Unemployed or in education / training

3.3.1 Primary analysis – Household information

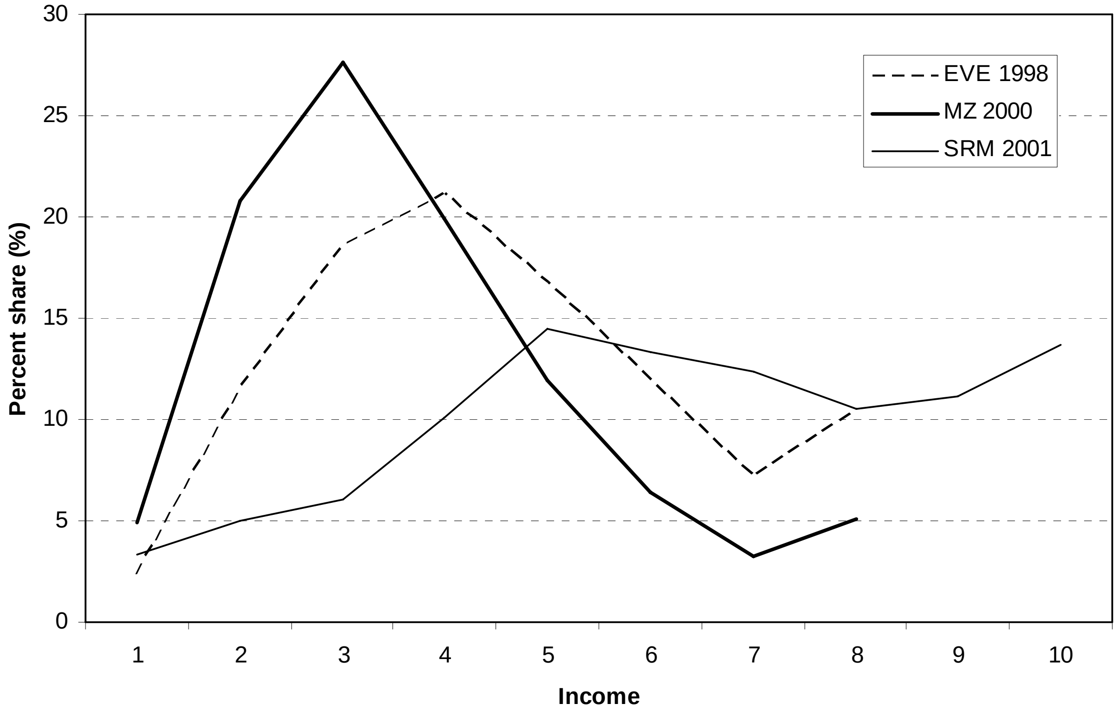
Household size distribution of the participated households helps in assessing the representation of the survey sample. Figure 3 shows the sample household size distributions of the three selected surveys. Household size in this analysis used is the number of permanently residing household members. A low share of single member households were observed in the SRM 2001 compared to the other surveys, which can be attributed to the survey context (leisure panel survey) and sampling unit (household). Apart from this, the household distributions of the three surveys are

Figure 3 Household size distributions



Monthly income is the common measure usually observed in any surveys. Figure 4 shows the income distributions of the three selected surveys. Due to the normalisation performed on the SRM 2001, monthly income has been over projected compared with the two surveys. Monthly income distribution of MZ 2000 is skewed to the left of the EVE 1998 income distribution. This can be addressed to the categorical observation in MZ 2000 and detailed income information collected in EVE 1998. The diverse distribution of monthly income information discourages to use the income variables in results interpretation.

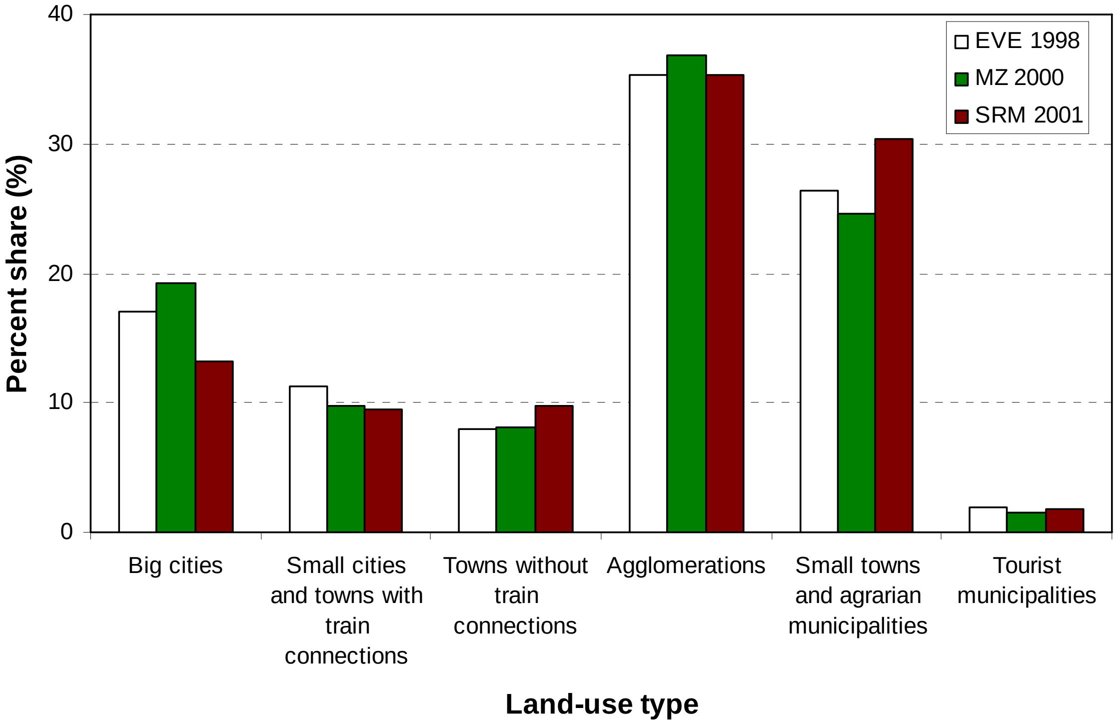
Figure 4 Income distributions



Source: ... (This and the previous line can be deleted, if not needed)

Household location land-use type: Figure 5 shows the transport-based household location land-use type distributions of the three selected surveys by land-use type. The distributions reveal that the country side people's active participation in longer version of the surveys compared to that of the people residing in cities and agglomerations.

Figure 5 Transport based household location land-use type distributions

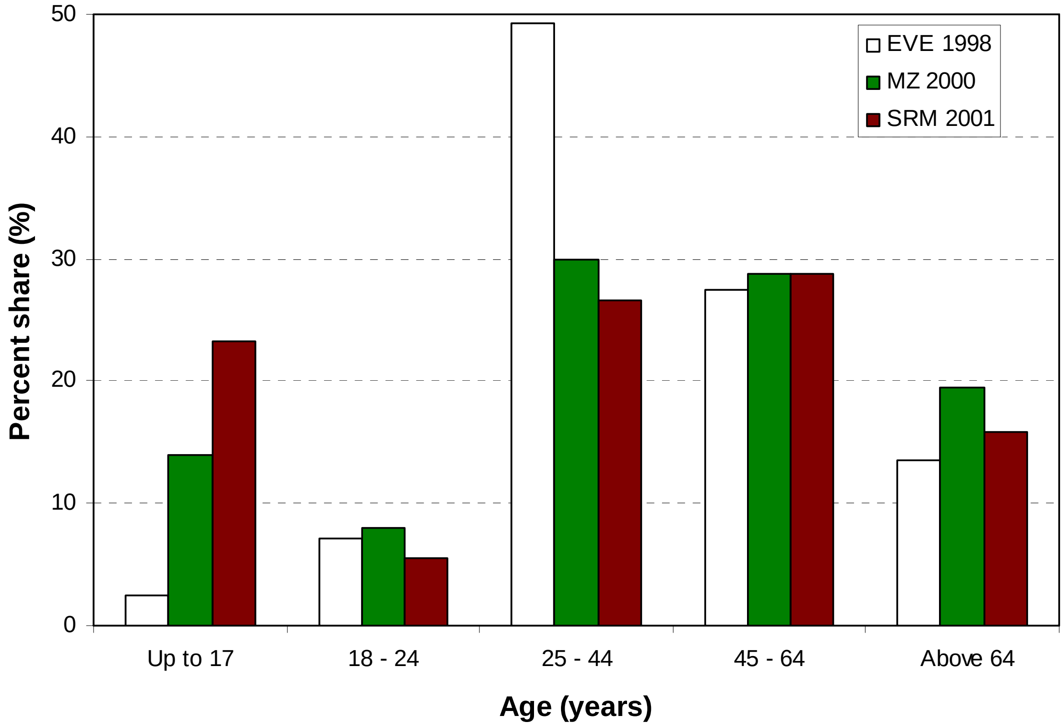


Source: ... (This and the previous line can be deleted, if not needed)

3.3.2 Primary analysis – Personal information

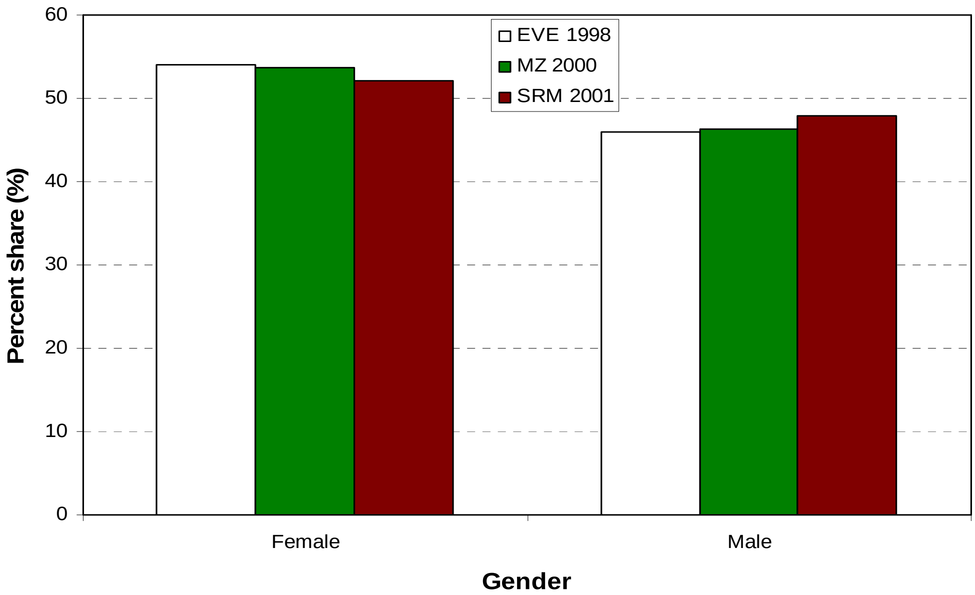
Age: Figure 6 shows the sample age distributions of the selected three surveys. Because of a higher age limit (15 years) of the respondents, EVE 1998 has a lower share of below 18 years group and a distinctly large share of adults aged between 25 to 44 years. The age distributions of MZ 2000 and SRM 2001 were found more or less consistent.

Figure 6 Age of the respondent



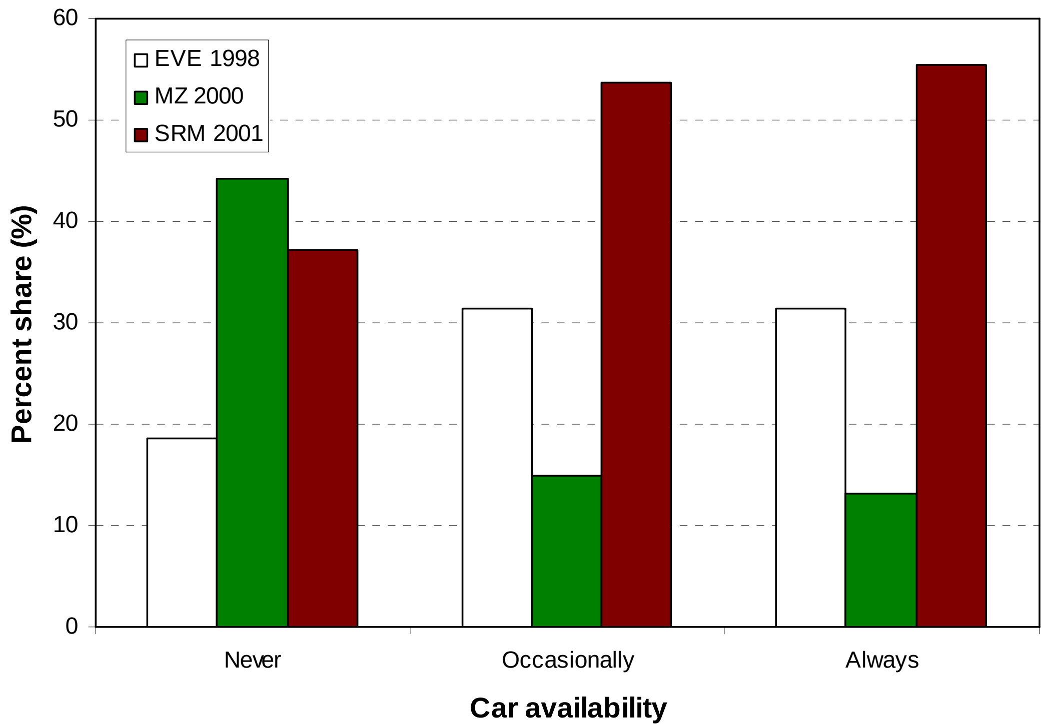
Gender: Figure 7 shows the gender shares of the three selected samples. Little to negligible variations among the gender distributions were noticed.

Figure 7 Title of the figure



Car availability: Figure 8 shows the car availability distributions. A distinct difference in car availability distributions of the three survey samples has been observed. It is noteworthy to mention the fact that the rural areas in Switzerland have a higher car ownership to that of cities. In other words, mathematically, car ownership is proportional to the land-use type in Switzerland. Higher share of car availability in SRM 2001 is due to the higher share of non-urban respondents.

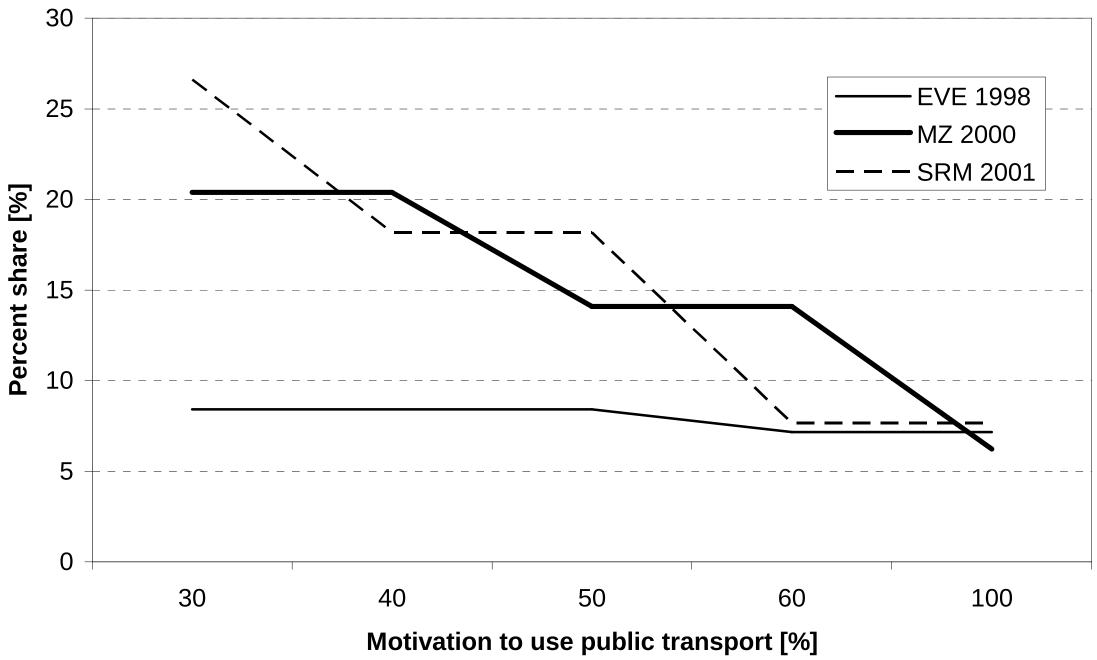
Figure 8 Distributions of car availability



Source: ... (This and the previous line can be deleted, if not needed)

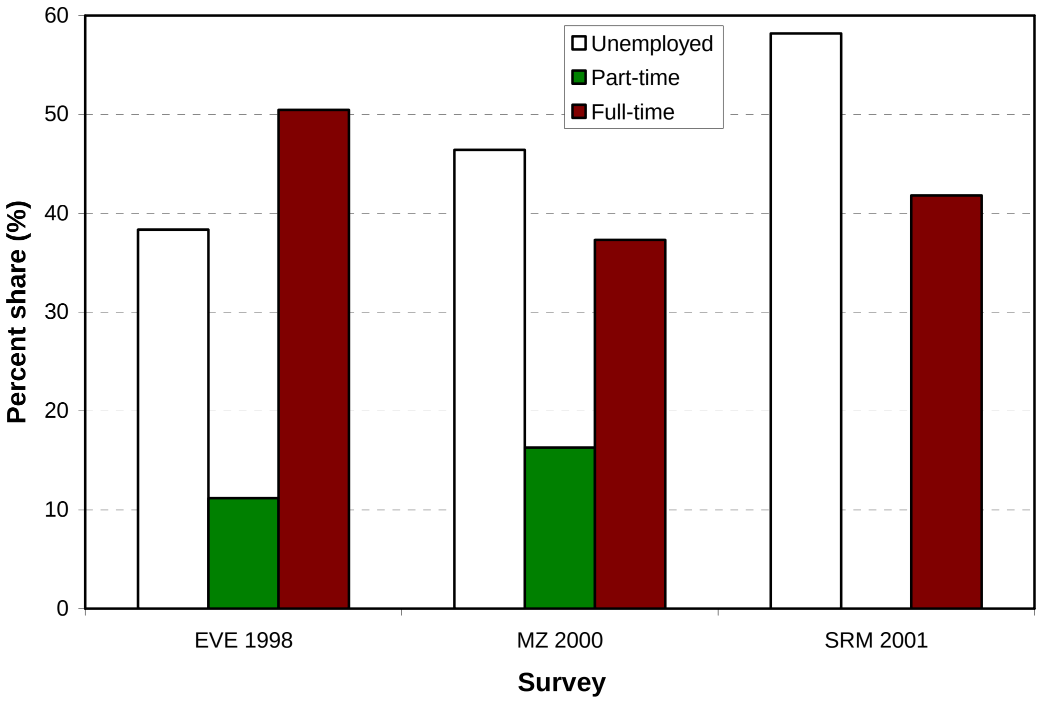
Public transport season tickets: Figure 9 shows the distributions of motivation to use public transport as a proxy to the mobility tools ownership in the form of public transport season tickets. Distributions for both the surveys MZ 2000 and SRM 2001 were found similar and motivation to use public transport was observed the least in EVE 1998.

Figure 8 Distributions of motivation to use public transport [%]



Working: Figure 10 shows the working status distributions of the three selected surveys. In the SRM 2001, working status was observed in two categories, while in the other two surveys an additional category in terms of “part-time” was observed. With equal number of observed categories, working status distributions of both the EVE 1998 and MZ 2000 were consistent. If the observed two categories share the missing category “part-time” proportionally, the working status distribution of the SRM 2001 is also consistent with that of two.

Figure 8 Distributions of working status



Source: ... (This and the previous line can be deleted, if not needed)

3.3.3 Primary analysis – Survey characteristics

Share of excursions in the reported travel information [%]: As this study focuses on combining the leisure travel information, the share of excursions in the reported travel information turns out to be crucial factor to assess the reliability of the reported information. The share of excursions in the reported travel information explains the extent of respondents thinking in terms of reporting excursions. Higher the share can be treated as more reliable information or information reported with more seriousness and vice versa. The share of excursions in the reported travel information for the three selected surveys is:

- EVE 1998: 16.67 %
- MZ 2000: 2.15%
- SRM 2001: 43.15%

Reporting period: Reporting period, duration Table 9 shows the reporting periods, fall-back period of the selected three surveys

Table 9 Trip types recorded by minimum trip duration, reporting period, and fall back period

Survey	Travel type	Duration	Reporting period	Fallback period
EVE 1998	Private same-day journeys*	1 day max.	2 weeks	-
	Short private journeys	1 – 3 nights	3 months	9 months
	Long private journeys	> 3 nights	6 months	6 months
	Business same-day journey	1 day max.	2 weeks	-
	Short business journey	1 – 3 nights	3 months	9 months
	Long business journeys	> 3 nights	6 months	6 months
MZ 2000	Trips*	1 day max.	1 day	-
	Long journeys	>1 night	3 months	-
	Long journeys (by air)	>1 night	12 months	-
	Long journeys (by air)	>1 night	5 years	-
SRM 2001	Excursions*	1 day max.	1 year	-
	Long journeys	>1 night	1 year	-

*: Excursions

3.4 Data preparation

The data preparation process, which consists of several sequential and individual steps, starts after finalising the variables of common information and leads to a final format that can be directly fed into the WinBUGS modelling structure. The input data in WinBUGS modelling environment is either S-PLUS format or rectangular format, while the former is most recommended. After identified the common variables, each variable was verified for the item non-response (or missing values). With the single-stage hot-deck imputation procedure using Solas 3.0, missing values for five variables were imputed. A simple macro in “Tcl/tk” was written to convert the data from plain ASCII format to S-PLUS format. The input data in S-PLUS format was carefully mined for each model.

4. MODELLING: VARIABLES, PARAMETERS AND MODEL FORM

4.1 Summary and general approach proposed

WinBUGS is the windows version of BUGS (Bayesian inference Using Gibbs Sampling), an open source flexible software for the Bayesian analysis of complex statistical models using Markov Chain Monte Carlo (MCMC) methods. The development of Markov chain Monte Carlo (MCMC) methods has made it possible to fir increasingly large classes of models with aim of exploring real world complexities of data (Gilks *et al.*, 1996). This ability naturally allows us to compare alternative model formulations. Taking advantage of this, a set of models were developed to combine leisure travel information from multiple sources.

This chapter deals with the modelling process of combining leisure travel information from multiple sources. Detailed description on the variables and

4.2 Common object observed in the data sources

As specified in the proposed framework explained earlier, selected data should be analysed to compare the data to obtain information in common among the multiple sources. “Excursion” was considered as the common object information to combine

4.3 Graphical models in WinBUGS

First step to analyse the travel demand in WinBUGS is construction of a directed graphical model. The graphical model represents all quantities as nodes in a directed graph and a node except its descendents, is independent of all other nodes. A special drawing tool *DoogleBUGS* has been developed for specifying graphical models, which uses a hyper-diagram approach to add extra information to the graph to give a complete model specification. Doodle consists of three elements: *Nodes*, *Plates* and *Edges*.

Nodes: Nodes represent quantities (variables and parameters). Nodes in the graph are of three types – *Stochastic*, *Deterministic* and *Constants*. Stochastic nodes represent variables that are given a distribution, and are denoted as ellipses in the graph. Stochastic nodes may be observed in which case they are data, or may be unobserved and hence be parameters, which may be unknown quantities underlying a model. Observations on an individual case that are unobserved say due to censoring, or simply missing data. Constant nodes are fixed by the design of the study. They are always founder nodes (i.e. do not have parents), and are denoted as rectangles in the graph. Constant nodes must be specified in a data file. Deterministic nodes are logical function of other nodes.

Plates: Plates represent repeated parts of the graph. In brief, all the elements within the plate

Edges: Edges are directed links between two nodes. Edges are two types – a solid arrow indicates a stochastic dependence while a hollow arrow indicates a logical function. An undirected dashed link may also be drawn to represent an upper and lower bound, but not used in any of the models.

The conditional independence assumptions represented by the graph mean that the full joint distribution of all quantities *V* has a simple factorisation in terms of the conditional distribution p(v | parents[v]) of each node given its parents, so that

$$p(V) = \prod p(v | parents[v])$$

The crucial idea is that only the parent-child distributions need to be provided in order to fully specify the model, and *WinBUGS* then sorts out the necessary sampling methods directly from the expressed structure.

As an alternative to the graph in terms of a Doodle representation, the model can be specified using the text-based *BUGS* language, headed by the model statement:

model {

Text-based description of graph in BUGS language

}

4.4 Variables

All the selected variables representing common information were classified in four distinct categories:

- Socio-demographics : *Age*, g*ender, household size*, *income*, and *working status*
- Mobility tool ownership: *Car availability* and *public transport season tickets*
- Survey characteristics: *Reporting period* and *percent share excursions of the reported travel*
- Land-use: *household location*

The categories and primary analysis of each of these variables can be found in the previous chapter. Missing values in the raw data of some of the variables were imputed using SOLAS software (Solas, 3.0). An attempt was made to read the imputed values as prior information. More details can be found in the model description. Table 10 shows the list of independent variable used in the random effect linear regression models combining the excursion data from multiple travel surveys. All these variables are constant in type, and are denoted by rectangles in the graphical models.

Table 10 Independent variables used in the model

Name	Label
age	Age of the person
gender	Gender of the respondent
hh_size	Size of the household
hhincome	Monthly household income
hhead_inc	Monthly household head's income
working	Working status
PT_tickets	Public transport season ticket
car	Car availability
per_inf	Percent share of excursions in the reported travel information
rep_prd	Duration of the reporting period (days)
tfc	Household location land-use type

4.5 Model Parameters

Various parameters have been used to construct linear regression models. All these parameters have been presented as stochastic nodes and are denoted by ellipses in the graphical models. To avoid any unnecessary confusion with the source and associated variable, the parameter name was constructed in two parts like "*a.b*". Where, "a" denotes the source i.e. survey name such as mz, eve, srm, and f (full), and "b" denotes the associated variable name. As a part of the model specification in WinBUGS, each stochastic variable's prior distribution specifications such as name of the distribution, mean and precision; must be defined. Table 11 shows the model parameters prior distributions specifications and the associated variable. A simple normal distribution was considered due to the computational advantage. All the linear regression models should be constructed using the variables and parameters.

Table 1 Model parameters prior distributions specifications and associated variables

Parameter	Distribution name	Mean	Precision	Associated variable
*.age	Normal	0.0	0.000001	*_age
*.gender	Normal	0.0	0.000001	*_gender
*.hh_size	Nomal	0.0	0.000001	*_hh_size
*.income	Normal	0.0	0.000001	*_income
*.working	Normal	0.0	0.000001	*_working
*.PT_tickets	Normal	0.0	0.000001	*_PT_tickets
*.car	Normal	0.0	0.000001	*_car
*.per_inf	Normal	0.0	0.000001	*_per_inf
*.rep_prd	Normal	0.0	0.000001	*_rep_prd
*.tfc	Normal	0.0	0.000001	*_tfc
*.trip	Gamma	0.001	0.001	*_trip

[*] to be replaced with the corresponding survey

Random effect regression models: The terms “random” and “fixed” are used in the context of ANOVA and regression models, and refer to a certain type of statistical model. Almost always, fixed effects regression or ANOVA are used for the situations rarely involving random effects analyses. A fixed effects ANOVA refers to assumptions about the independent variable and the error distribution for the variable. An experimental design is the easiest example for illustrating the principal. Usually, the fixed effect regression model is applied to generalise the results to experimental values used in the study. However, if we want to make inferences beyond the particular values of the independent variable used in the study, a random effects model is used. In this study, we would like to make inferences beyond the three sources used, so the surveys are conceptualized as surveys randomly drawn from a larger universe of possible surveys and the inferences are made to a larger universe of leisure travel and range of abstractness values. Such a generalization is more of an inferential leap, and, consequently, the random effects model is less powerful.

Fixed and Random Coefficients in Multi-level Regression: The random vs. fixed distinction for variables and effects is important in multi-level regression. In multi-level regression models, both level-1 and level-2 *predictors* are assumed to be fixed. However, level-1 *intercepts and slopes* are typically assumed to vary randomly across groups. Because of the assumptions about their error distributions, we call their variances as “random coefficients.”

To understand the variations from the multiple sources as random effects, a set of random-effect parameters or “random coefficients” have been defined. All the random effect parameters used in various models of this study are stochastic. Table 12 shows the random-effect parameters prior distribution specifications along with the associated variables. These random-effect parameters are of two levels, 1st order parameters called as mean parameters and 2nd order parameters called as variance parameters. For a computational advantage, all the 1st order random-effect parameters priors are normally distributed. To study the prior’s informative property, a set of prior distributions with different specifications have been specified for the variance parameters (Gelman, 2005). The prior distributions for variance parameters considered are:

- Gamma distribution with mean = 0.001 and precision = 0.001
- Gamma distribution with mean = 1.0 and precision = 1.0
- Uniform distribution with mean = 0.0 and precision = 1000

Table 12 Specifications of the random-effect parameters prior distributions

Parameter	Distribution	Mean	Precision	Associated variable
age	Normal	0.0	0.000001	*_age
age_var	Gamma	0.001	0.001	*_age
gender	Normal	0.0	0.000001	*_gender
gender_var	Gamma	0.001	0.001	*_gender
hhsize	Nomal	0.0	0.00001	*_hh_size
hhsize_var	Gamma	.001	0.001	*_hh_size
income	Normal	0.0	0.000001	*_income
income_var	Gamma	0.001	0.001	*_income
working	Normal	0.0	0.000001	*_working
working_var	Gamma	0.001	0.001	*_working
PT_tickets	Normal	0.0	0.000001	*_PT_tickets
pt_var	Gamma	0.001	0.001	*_PT_tickets
car	Normal	0.0	0.000001	*_car
car_var	Gamma	0.001	0.001	*_car
per_inf	Normal	0.0	0.000001	*_per_inf
per_inf_var	Gamma	0.001	0.001	*_per_inf
rep_prd	Normal	0.0	0.000001	*_rep_prd
rep_prd_var	Gamma	0.001	0.001	*_rep_prd
tfc	Normal	0.0	0.000001	*_tfc
tfc_var	Gamma	0.001	0.001	*_tfc
trip_var	Gamma	0.001	0.001	*_trip

[*] to be replaced with the corresponding survey

The statistics shown in the Table 12 are the first case. A simple analysis on the distribution type against the strength of the informative prior distribution is aimed in the post-modelling phase.

4.6 Model form

The random effect regression model to estimate the leisure travel from multiple sources was developed in four stages:

Stage I: Socio-demographics (Socio)

Stage II: Socio-demographics and Mobility tools (So-Mo)

Stage III: Socio-demographics, Mobility tools and Survey characteristics (Wotfc)

Stage IV: Socio-demographics, Mobility tools, Survey characteristics and Land-use type (Complete)

WinBUGS strongly recommend to

Excursions = F{ Socio-demographics, Mobility tools, Survey characteristics, Regional traffic type}

4.6.1 *Stage I* – Socio model

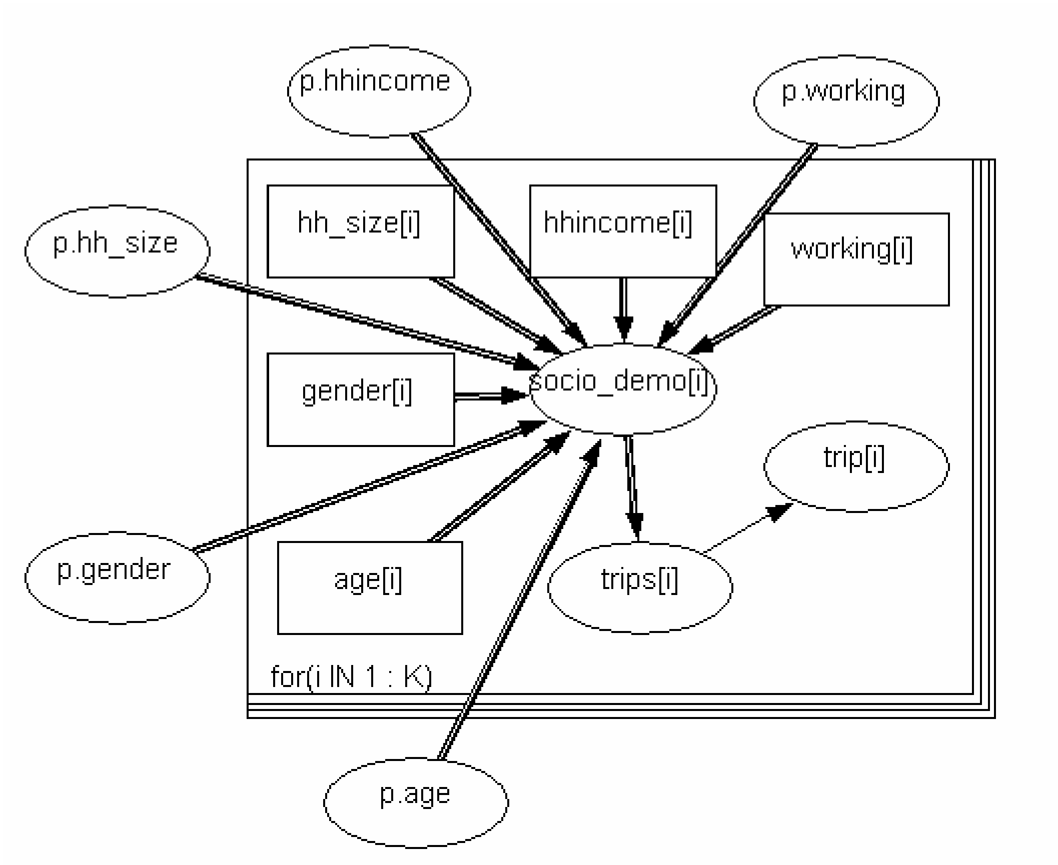
This is the first stage of the model development in the process of combining the leisure travel information from multiple sources. Specific assumption pertaining to this model is only socio-demographic characteristics of the travel survey influence the travel demand and other factors influence though present, is negligible. Socio is a linear regression model with least number of variables and parameters. Model form of the "Socio" model is explained the following equations:

Excursions = *F(Socio-demographics)*

Excursions = *f(Age, Gender, Household size, Income, Working status)*

A multi-level linear regression model with Bayesian inference was implemented to analyse the travel demand estimation processes. Using "Doodle" module in WinBUGS, a graphical model was developed and is shown in Figure 11.

Figure 11 Graphical version of the *Socio* model with single source



The graphical model shown in the Figure 11 is a fixed effect linear regression model. WinBUGS facilitates to extract the model code in text format and the model code of the socio model with single data source is shown in Figure 12.

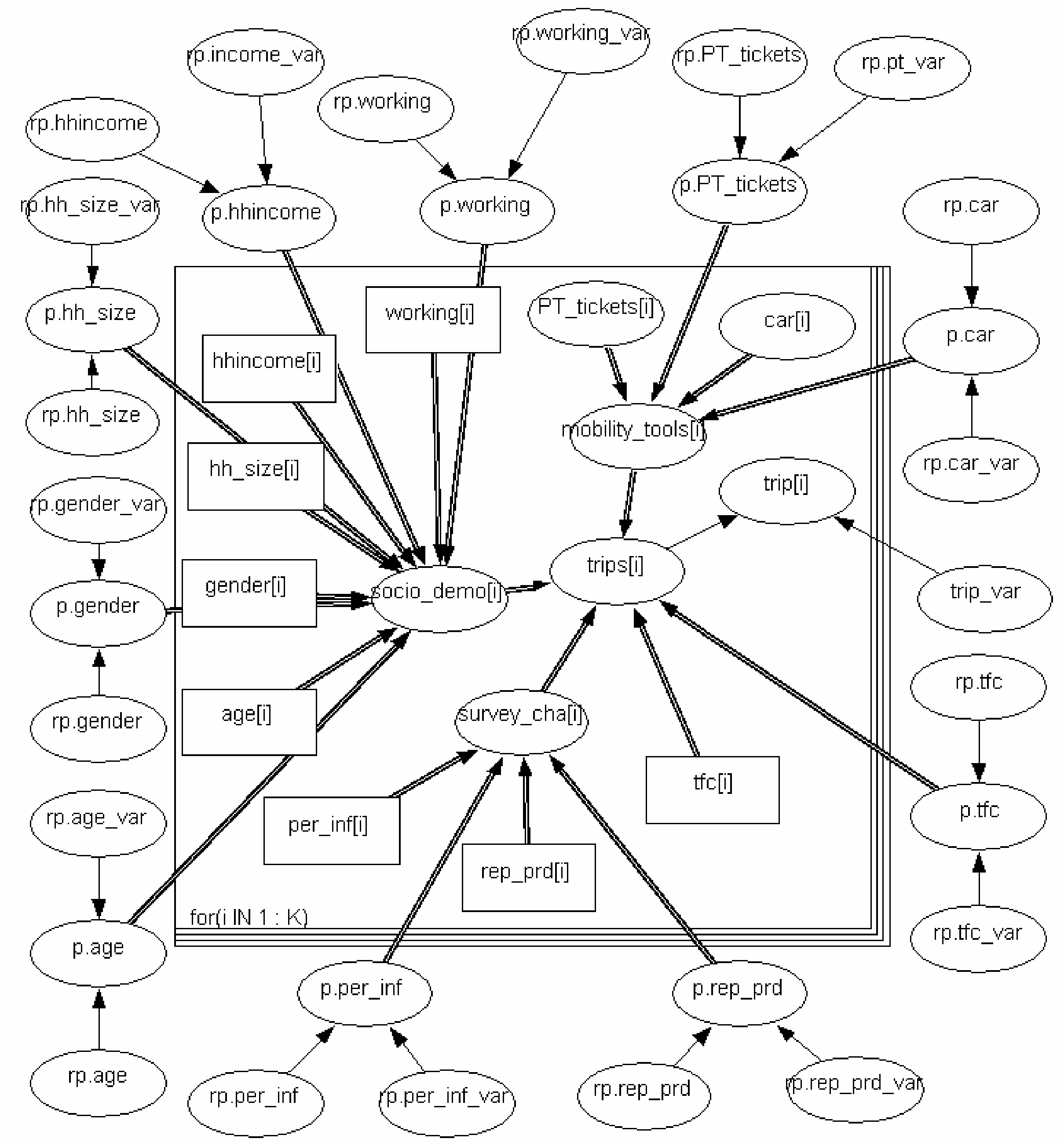
Figure 12 The *Socio* model code in *BUGS* language for a single source

```
model;
{
/* Defining the model parameters*/
  p.age ~ dnorm( 0.0,1.0E-6)
  p.gender ~ dnorm( 0.0,1.0E-6)
  p.hh_size ~ dnorm( 0.0,1.0E-6)
  p.hhincome ~ dnorm( 0.0,1.0E-6)
  p.working ~ dnorm( 0.0,1.0E-6)

/* Linear regression equation repeated for K times */
  for( i in 1 : K ) {
    socio_demo[i] <- age[i] * p.age + gender[i] * p.gender + hh_size[i] * p.hh_size +
                        hhincome[i] * p.hhincome + working[i] * p.working
  }
/* Factors that influence travel demand */
  for( i in 1 : K ) {
    trips[i] <- socio_demo[i]
  }
/* Travel demand estimation definition*/
  for( i in 1 : K ) {
    trip[i] ~ dnorm(trips[i],1.0E-6)
  }
}
```

A set of random coefficients have been added to the fixed effect linear regression model shown in the Figure 11 and the resulted random-effect linear regression model is shown in the Figure 13.

Figure 13 Random effect regression model (*Complete*) with single data source



4.6.2 *Stage II* – Socio-Mobility model (SOMO)

This is the second stage of the model development in the process of combining the leisure travel information from multiple sources. Specific assumption pertaining to this model is – only socio-demographic characteristics and mobility tools observed in travel survey influence the travel demand and other factors influence though present, is negligible. SoMo is also a linear regression model with number of variables and parameters. Model form of the "SoMo" model is explained in the following equations:

Excursions = *F(Socio-demographics, Mobility tools ownership)*

Excursions = *f(Age, Gender, Household size, Income, Working status) + f(Car availability, Public transport season tickets)*

A random effect linear regression model with Bayesian inference was implemented to analyse the travel demand estimation processes. Using "Doodle" module in WinBUGS, a graphical model was developed and the code is shown in Appendix A.

4.6.3 *Stage III* – Socio-Demographics, Mobility tools and Survey characteristics (Wotfc)

In this Stage III model, only land-use type of home location is excluded from the full model. "Wotfc" model is explained in the following equations:

Excursions = *F(Socio-demographics, Mobility tool ownership, Survey characteristics)*

Excursions = *f(Age, Gender, Household size, Income, Working status) + f(Car availability, Mobility tools ownership) + f(Percent share of excursions in reported travel, Reporting period)*

A random effect liner regression model with Bayesian inference was implemented to analyse the travel demand estimation processes. Using "Doodle" module in WinBUGS, a graphical model was developed and the code is shown in Appendix A.

4.6.4 Stage IV – Complete model

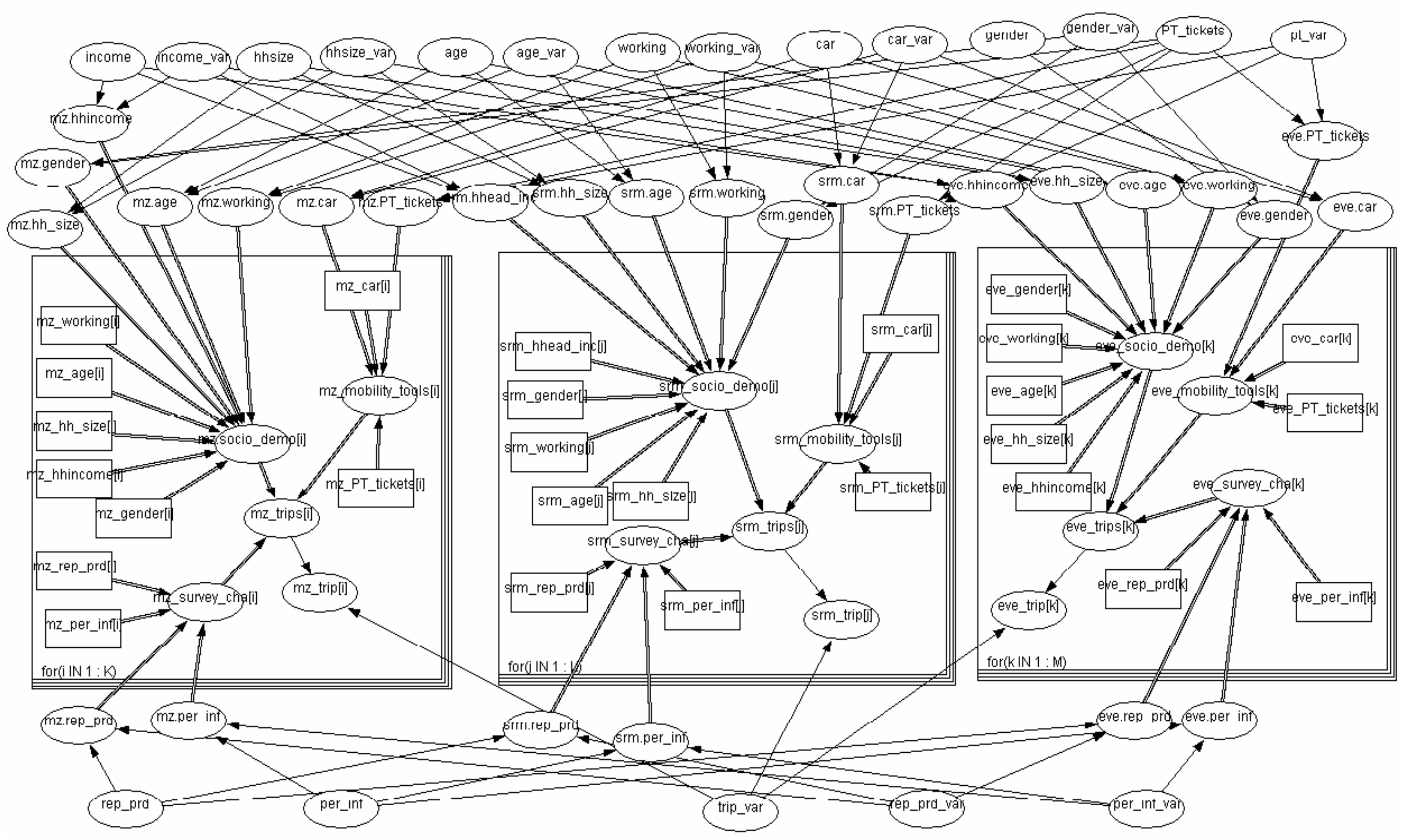
Stage IV is the final stage of the model development to combine leisure travel information from multiple sources. Important assumption of the complete model is all the four distinct groups of variables namely, Socio-demographics, Mobility tools ownership, Survey characteristics, and Land-use, influence the travel demand. "Complete" model is explained in the following equations:

Excursions = *F(Socio-demographics, Mobility tools ownership, Survey characteristics, Land-use)*

Excursions = *f(Age, Gender, Household size, Income, Working status) + f(Car availability, Mobility tools ownership) + f(Percent share of excursions in reported travel, Reporting period) + f(Land-use type of the household location)*

The graphical version of the random effect linear regression model with Bayesian inference implemented to analyse the travel demand estimation process in stage IV is shown in Figure 14. The model code is scripted in Appendix A.

Figure 14 Graphical model to integrate the excursions from multiple sources



In addition to the four stage model, another simple model named *Combined* was also developed.

4.6.5 Combined model

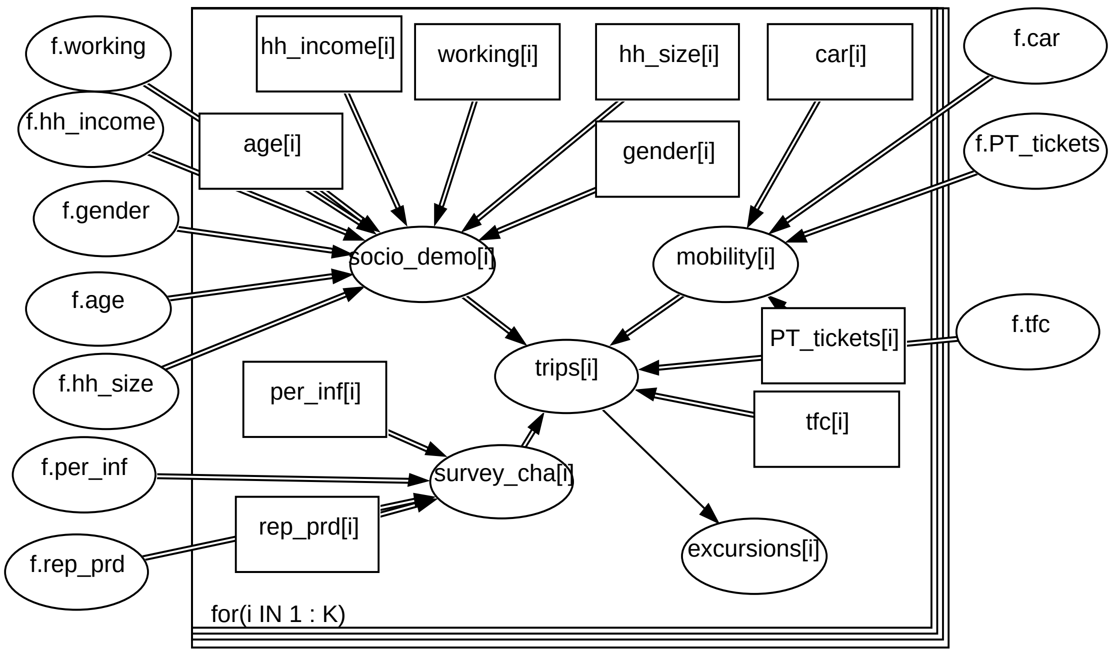
The explained four stage model assumed that the selected surveys have collected the information independently. In the *Combined* model the common information was treated as the information observed in three surveys simultaneously conducted with different reporting periods. “Combined” model can be explained with the following equations:

Excursions = *F(Socio-demographics, Mobility tools ownership, Survey characteristics, Land-use)*

Excursions = *f(Age, Gender, Household size, Income, Working status) + f(Car availability, Mobility tools ownership) + f(Percent share of excursions in reported travel, Reporting period) + f(Land-use type of the household location)*

The random effects cannot be measured for this single source model. However, the graphical version of the linear regression model with Bayesian inference implemented to analyse the travel demand estimation process is shown in Figure 15.

Figure 15 Graphical version of the *Combined* model with single source



4.7 Modelling the travel demand estimation

Modelling process involves sequential testing of the four defined random effect linear regression models for the data from three selected surveys in WinBUGS environment. This study followed the standard WinBUGS model run procedure and is explained in the following sequential steps:

Step 1: Specify the graphical model using "Doodle" editor. Validate the model for its consistency with the BUGS language code and extract the model code.

Step 2: Highlight the word "model" in the code and check the model, which parses the BUGS language description of the statistical model. If a syntax error is detected the cursor is placed where the error was found and the description of the error is given on the status line.

Step 3: Loading the data can be performed in two ways: 1) if the data is in a separate document, then open the document and use *load data* command. 2) if the data is a part of the document, the first character of the data (here *list* as the data is in S-Plus format) should be highlighted and use *load data* command.

Step 4: Specify the number of chains i.e. sets of samples to simulate. Since simulating with multiple chains is one to check the convergence of the MCMC simulations, it is generally preferred to simulate with multiple chains for simple models. If the model is fairly complex, a pilot run using a single chain to check the model compilation. Once the model is running smoothly with a single chain, re-run it with multiple chains. This two stage model run process was successfully implemented for all the models and two chains were used.

Step 5: After defining the number of chains, compile the model with a single click on "compile" button in the model specification tool window. A message saying "model compiled" should appear in the bottom left of the WinBUGS program. This sets up the internal data structures and chooses the specific MCMC updating algorithms to be used by WinBUGS for the compiled model.

Step 6: The MCMC sampler must be given some initial valued for each stochastic node. These can be arbitrary values, although in practice, convergence can be poor if wildly inappropriate values are chosen. A different set of initial values for each chain are required. There is no need to provide a list of initial values for every parameter in the model, but can also be generated. WinBUGS generates initial values for any stochastic parameter not already initialised. The initial value generation is performed by forward sampling from the prior distribution for each parameter. All the initial values for the stochastic parameters in the models were provided in this study and care was taken to avoid wildly inappropriate values. To load the initial values, highlight the *list* at the beginning of the set of initial values and click the *load inits* button. A message saying "initial values loaded: model initialised" should appear in the bottom left of the WinBUGS program window.

Step 7: Now the initialised model is ready to start running the simulation. Before doing so, it is essential to set some monitors to store the sampled values for selected parameters. The monitors for the parameters can be set in samples window, but the manual process is quite cumbersome. The convenient way to set the monitors for the parameters is using the script, which can be re-used.

Step 8: To run the simulation, the number of updates (iterations of the simulation) must be specified in the *Update* window. When updates are finished, the message “updates took *** s” will appear in the bottom left of the WinBUGS window program. Checking for simulation convergence is an important step in which the model parameters are individually tested for convergence. This convergence check is performed with the previously set monitors for the parameters. For models with many parameters, it is performed only on the selected parameters. For all the models in this study convergence check was performed only after 10’000 iterations.

Step 9: Once confirmed that the simulation has converged, it is necessary to run some further updates to obtain a sample from the posterior distribution. The outcome of this modelling process is the summaries of these posterior distributions such as statistics, kernel densities, pair-wise correlations, etc. Finally, using the model output analysis on model fit and complexity should be performed.

Modelling the leisure travel estimation is summarised as:

- Four incremental models were constructed to understand the model development process
- An additional model assuming that the common information was derived from a single source multi-period and multi-mode travel survey.
- An exclusive script for each model was created and used to simulate the model.
- Missing values in the observed variables were imputed and treated in two ways in modelling process – imputed values were fed into the model as initial values , and imputed values were assumed as equivalent observed values and treated as observed data
- Two simulation chains were used and each model was updated for a million iterations.
- Output in the form of summary statistics of the posterior distributions and

5. POST-MODELLING: RESULTS AND INTERPRETATION

Post-modelling phase includes analysing the results obtained from modelling, interpretation and drawing vital conclusions in the lines of the study objectives. This chapter covers the first two only i.e. analysing the results and interpretation of the same.

Results obtained from modelling are nothing but output from WinBUGS model. Output from WinBUGS model considered is inferences of the three elements, namely Samples, Correlations, and DIC. Samples inferences extracted for all the models in this study are – stats (summary statistics of the posterior distributions), history (plot the complete trace of the variable), quantiles (running mean with running 95% confidence intervals) and densities (plot of smoothened kernel density). Correlations inference extracted is correlation coefficients among the selected variables. The DIC inference element calculates a set of statistics used in model complexity and fit analysis.

For a better understanding and consistency, descriptions of both results and interpretation have been meaningfully weaved. Two of the three output elements, posterior distributions and pair-wise correlations are described first followed by the description on estimation of the excursions using posterior distributions.

5.1 Posterior distributions

Posteriors are the main outcome of any Bayesian analysis, often called as Bayesian inferences. Not surprisingly, posteriors occupy central place in WinBUGS output. After updating the random effect regression model for a million times, the posterior samples were summarised in two ways:

- Numerical summary – involves calculation of summary statistics such as the mean, variance and quantiles of the sample.

- Graphical summary – covers plotting kernel densities, historical summary, and tracing the posterior samples.

Summary of the posterior distribution samples were extracted with the *sample monitor tool* window by selecting the previously specified monitors for the parameters and the iterations range. The sample monitor tool facilitates to extract the summaries for both within as well as across simulation chains. Table 13 shows the summary statistics of the posterior distribution samples of *Socio* model. Two important statistics, mean and standard deviation, are listed in the table where are the percentiles are excluded. The quantity MC error is the Monte Carlo standard error of the mean. Start represents the starting update and sample is the posterior sample size, i.e. difference between start and end updates specified in summary monitor tool. In addition to these, percentiles are also a part of the summary statistics. Summary statistics for the three other models are shown in Appendix A.X, A.Y, and A.Z.

Table 13 Summary statistics of *Socio* model posterior distribution sample

Parameter	Mean	Std. Dev	MC error	Start	Sample
MZ 2000					
mz.age	0.00335	0.004222	8.88E-06	1000	999001
mz.gender	0.007248	0.0172	2.77E-05	1000	999001
mz.hh_size	0.005473	0.005557	1.45E-05	1000	999001
mz.hhincome	0.002525	0.005212	1.56E-05	1000	999001
mz.working	-0.00009	0.01976	3.50E-05	1000	999001
SRM 2001					
srm.age	0.6172	0.0111	2.47E-05	1000	999001
srm.gender	-0.02111	0.03897	6.63E-05	1000	999001
srm.hh_size	0.1172	0.01314	3.75E-05	1000	999001
srm.hhead_inc	0.1008	0.007698	2.41E-05	1000	999001
srm.working	-0.7615	0.04393	7.17E-05	1000	999001
EVE 1998					
eve.age	0.0673	0.01325	3.41E-05	1000	999001
eve.gender	0.07509	0.04508	8.88E-05	1000	999001
eve.hh_size	0.02086	0.01637	4.31E-05	1000	999001
eve.hhincome	0.02474	0.01219	3.93E-05	1000	999001
eve.working	-0.1104	0.04977	1.04E-04	1000	999001
Joint					
age	0.228	0.7485	7.82E-04	1000	999001
gender	0.02052	0.1465	1.32E-04	1000	999001
hhsize	0.04772	0.184	1.83E-04	1000	999001
income	0.04298	0.1178	1.16E-04	1000	999001
working	-0.2898	0.8105	8.12E-04	1000	999001

5.2 Correlation matrices

WinBUGS facilitates to calculate pair-wise correlations for all the previously set monitors for various stochastic parameters through the *Correlation tool* window. Table 14 shows the correlation matrix for the Stage I model (*Socio*). One of the major advantages in producing these correlation matrices is to have a platform to understand the interaction among various model parameters. But this study used correlation matrices for estimating the travel demand using posteriors distribution samples, which is covered in the next section. Correlation matrices for the other three models are shown in Appendix A.9.1, A.9.2, and A.9.3.

Table 14 Correlation matrix of the *Stage I* model

	age	gender	hhsize	income	working
EVE 1998					
age	1	-0.1770	-0.1659	-0.4923	-0.0493
gender	-0.1770	1	-0.1019	0.0290	-0.3663
hhsize	-0.1659	-0.1019	1	-0.5618	0.0831
income	-0.4923	0.0290	-0.5618	1	-0.3618
working	-0.0493	-0.3663	0.0831	-0.3618	1
MZ 2000					
age	1	-0.1235	-0.0746	-0.4244	-0.1291
gender	-0.1235	1	-0.1772	-0.0471	-0.2823
hhsize	-0.0746	-0.1772	1	-0.6348	0.1322
income	-0.4244	-0.0471	-0.6348	1	-0.3459
working	-0.1291	-0.2823	0.1322	-0.3459	1
SRM 2001					
age	1	-0.1000	-0.0628	-0.4320	-0.2348
gender	-0.1000	1	-0.2115	-0.0563	-0.1958
hhsize	-0.0628	-0.2115	1	-0.6656	0.0966
income	-0.4320	-0.0563	-0.6656	1	-0.2186
working	-0.2348	-0.1958	0.0966	-0.2186	1
Joint					
age	1	-0.0003	0.0010	-0.0009	0.0009
gender	-0.0003	1	-0.0016	-0.0007	-0.0009
hhsize	0.0010	-0.0016	1	-0.0011	0.0003
income	-0.0009	-0.0007	-0.0011	1	0.0002
working	0.0009	-0.0009	0.0003	0.0002	1

5.3 Estimating the excursions using posterior distributions

The motive to combine the similar information from different sources is neither to assess the information quality of each individual source nor to estimate the elasticities, but to analyse the combined effect on the theme of interest. Here leisure travel in the form of excursions was observed from three independent travel surveys within a small time range (1998-2001). A random effect linear regression model with Bayesian inference was constructed and the join posterior distribution samples for all stochastic parameters were obtained from the modelling process. A modified excursions distribution was estimated using four elements: correlation matrix, summary of posterior distribution samples of the model parameters, linear regression model and the observed travel information. The sequential steps of this estimation process are:

- Using the correlation matrix, a set of normally distributed correlated random numbers with zero mean and one as standard deviation for all parameters were generated.
- These random generated distributions were adjusted to the corresponding parameters. This correlated random distributions generation process was performed in MATLAB and Figure 16 shows the script used for this process.
- A sub-set equivalent of survey sample size was randomly selected. Excursions were calculated using the linear regression relationship defined in the model. For the randomly selected posterior distributions based random generated parameters and observed travel information, modified excursions for each respondent were calculated.

Figure 16 MALAB code for correlated random parameters generation

```
%Create dependent random variables
%This function generates N independent simulation of M dependend r.v.
%Input parameters
%M       - number of dependend random variables ;
%N       - number of simulations;
M=9;
N=2000000;

%Correlation matrix
Correlation_Matrix=
[1  0.013   0.524   0.184   0.041   -0.170  -0.517  -0.620  0.022   0.093
0.013   1   -0.086  -0.018  -0.192  0.018   -0.021  -0.067  0.000   0.013
0.524   -0.086  1   -0.094  0.061   0.020   -0.045  -0.528  -0.027  -0.121
0.184   -0.018  -0.094  1   -0.137  -0.096  -0.184  -0.348  0.013   0.101
0.041   -0.192  0.061   -0.137  1   -0.053  -0.328  -0.010  0.000   -0.005
-0.170  0.018   0.020   -0.096  -0.053  1   0.212   -0.075  0.019   0.094
-0.517  -0.021  -0.045  -0.184  -0.328  0.212   1   0.133   -0.008  -0.109
-0.620  -0.067  -0.528  -0.348  -0.010  -0.075  0.133   1   -0.535  -0.443
0.022   0.000   -0.027  0.013   0.000   0.019   -0.008  -0.535  1   0.191
0.093   0.013   -0.121  0.101   -0.005  0.094   -0.109  -0.443  0.191   1
]

%create correlation Matrix
Number_CM=zeros(1,M); %Number of Correlated Members

%Creates multivariant normal ditribution rnadom variables-also multivariant
Z=mvnrnd(Number_CM,Correlation_Matrix,N);

% Assign names to the correlated random parameters
Age = Z(:,1);
Gender = Z(:,2);
Hhsize = Z(:,3);
```

```
Income = Z(:,4);
Working = Z(:,5);
PT_tickets = Z(:,6);
Car = Z(:,7);
Per_inf = Z(:,8);
Rep_prd = Z(:,9);

% Adjust the random parameters to the summary of posterior distributions
complete_Age = (Age * Std. Dev) + mean;

% Array of correlated random generated parameters
full_eve = [eve_age eve_gender eve_hhsize eve_income eve_working
eve_pttickets eve_car eve_perinf eve_repprd eve_tfc];

% save to a file
save('C:\..\full_eve.dat', 'full_eve', '-ASCII');
```

Posterior distribution based excursions certainly differ from that of observed, and also depends on the modelling process that produced the posteriors. This study aimed to analyse the differences between observed and estimated excursions distributions for various models. Excursions were estimated using the regression models implemented to estimate the posterior distribution. For variables, observed data from respective surveys were used. For parameters, a sample equivalent to the survey sample was randomly selected from the random generated posterior distribution of monitors. The regression model forms used in the excursions estimation are:

Stage I – Socio

Excursions = *f*(age, gender, household size, monthly income, working status)

Stage II – So-Mo

Excursions = *f*(age, gender, household size, monthly income, working status, car availability, public transport season tickets)

Stage III – Wotfc

Excursions = *f*(age, gender, household size, monthly income, working status, car availability, public transport season tickets, share of excursion in the reported travel, reporting period)

Stage IV – Complete

Excursion = *f*(age, gender, household size, monthly income, working status, car availability, public transport season tickets, share of excursions in the reported travel, reporting period, household location land-use type)

In principle travel frequency is a non-negative term. But the range of estimated excursions was found extended into negatives, which is absolutely illogical from transport point of view. A simple counter analysis was performed to mitigate this conspiracy. This attempt involves shifting of the distribution from below-zero to positive range while maintaining the statistics. The iterative shifting process has the following three steps:

- Calculate the mean of the first iteration (M_0). Adjust all negatives to zero.
- Calculate the mean of the series with adjusted values (M_n)
- $X_{n+1} = X_n + [M_n – M_0]$

Table 15 shows the statistical summary of estimated excursions using posterior distributions.

Table 15 Statistical summary of the estimated excursions using posterior distributions

Model type	Component	No of cases	Minimum	Maximum	Mean	Std. Dev.	AAE*
Stage I	EVE 1998	4430	0.0063	0.8028	0.3671	0.1092	9.5434
	MZ 2000	29407	-0.0614	0.1577	0.0384	0.0211	14.0166
	SRM 2001	5004	-6.3442	4.5783	2.5626	0.8261	3.4317
	Combined	38841	-11.7034	15.2530	0.9829	2.8275	262.1869
Stage II	EVE 1998	4430	-0.0654	1.6097	0.3697	0.2879	9.6124
	MZ 2000	29407	-0.1010	0.1879	0.0386	0.0250	14.0729
	SRM 2001	5004	-6.2030	5.3441	2.5702	0.8674	2.5702
	Combined	38841	-12.8225	17.5018	0.9685	2.9207	266.8487
Stage III	EVE 1998	4430	-0.2647	1.7091	0.3228	0.3010	8.3933
	MZ 2000	29407	-0.1104	0.1893	0.0384	0.0258	14.0109
	SRM 2001	5004	-9.9344	7.0635	2.5567	1.5297	2.5567
	Combined	38841	-485.1690	567.3544	4.5719	48.8656	5.0063
Stage IV	EVE 1998	4430	-1.0741	2.1486	0.2934	0.4340	7.6295
	MZ 2000	29407	-0.0879	0.1833	0.0385	0.0277	14.0652
	SRM 2001	5004	-7.3413	5.4938	2.6199	1.1009	2.6199
	Combined	38841	-1134.518	1413.2192	6.2358	110.6160	6.6710
Cmb+	Combined+	38841	-31838.70	38351.7	33.174	4656.05	33.174

*: Average annual excursions +: Combined model

Statistics (Mean, Minimum and Maximum) shown in the Table XX are before eliminating the negative values. After adjusting the excursions to non-negative, average annual excursions were calculated by linearly extrapolating the average estimated excursions for the reporting period.

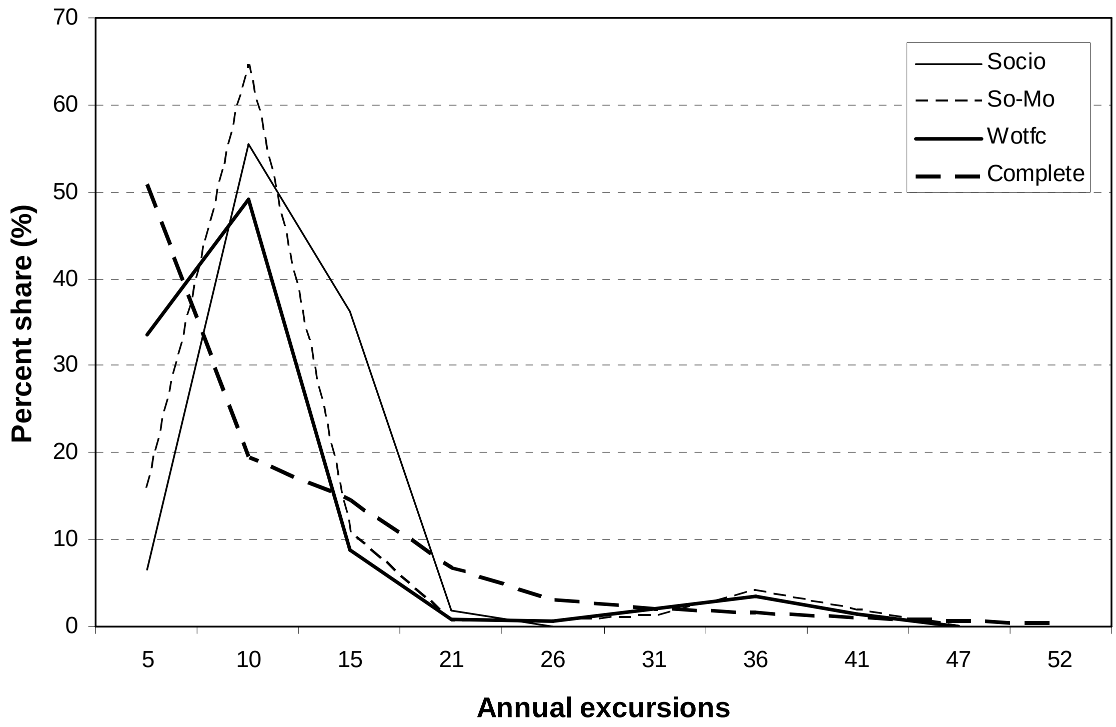
Estimated average annual excursions for individual surveys were found inversely proportional to the reporting periods of the surveys. But variations in the average annual excursions between different models were very small to negligible. The most interesting issue that need to be focussed is the variations of annual average excursions of the combined to that's of individual surveys. The combined average annual excursions estimated vary between different models. Specifically in the first two stages, average annual excursions were very high. This may be partly attached to the model redundancy. With the addition of survey characteristics, a sudden drop in the average excursions was observed. This tempting inference is sufficiently supported by the stage IV model, in which the average excursions estimated were close to that of Stage III model average. An important inference of this analysis is average excursions were largely influenced by the survey characteristics to that of respondent's individual characteristics.

Excursions for the EVE 1998 were estimated using the random generated

5.3.1 Travel module of "Household income and expenditure survey 1998"

Figure 17 shows the distributions of estimated excursions for various models. The a

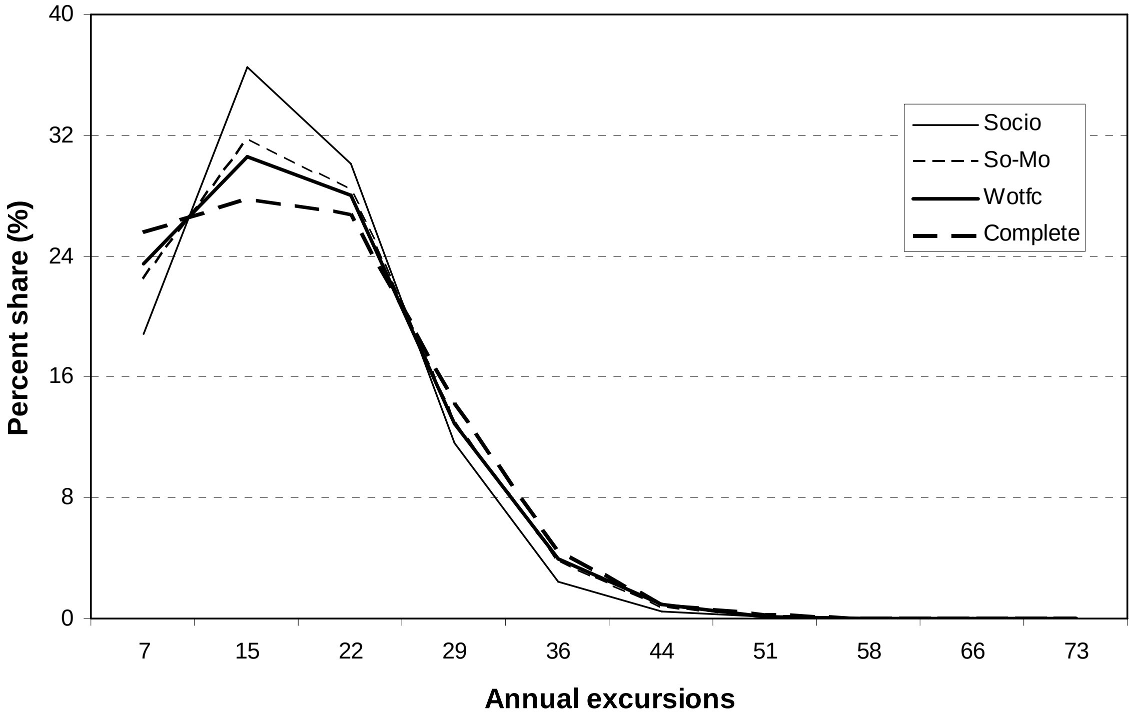
Figure 17 Distributions of estimated average annual excursions (EVE 1998)



5.3.2 Microcensus 2000

Figure 18 shows the distribution of linearly projected annual excursions estimated using the linear regression models and their respective randomly selected random generated parameter's posteriors.

Figure 18 Distributions of estimated average annual excursions (MZ 2000)

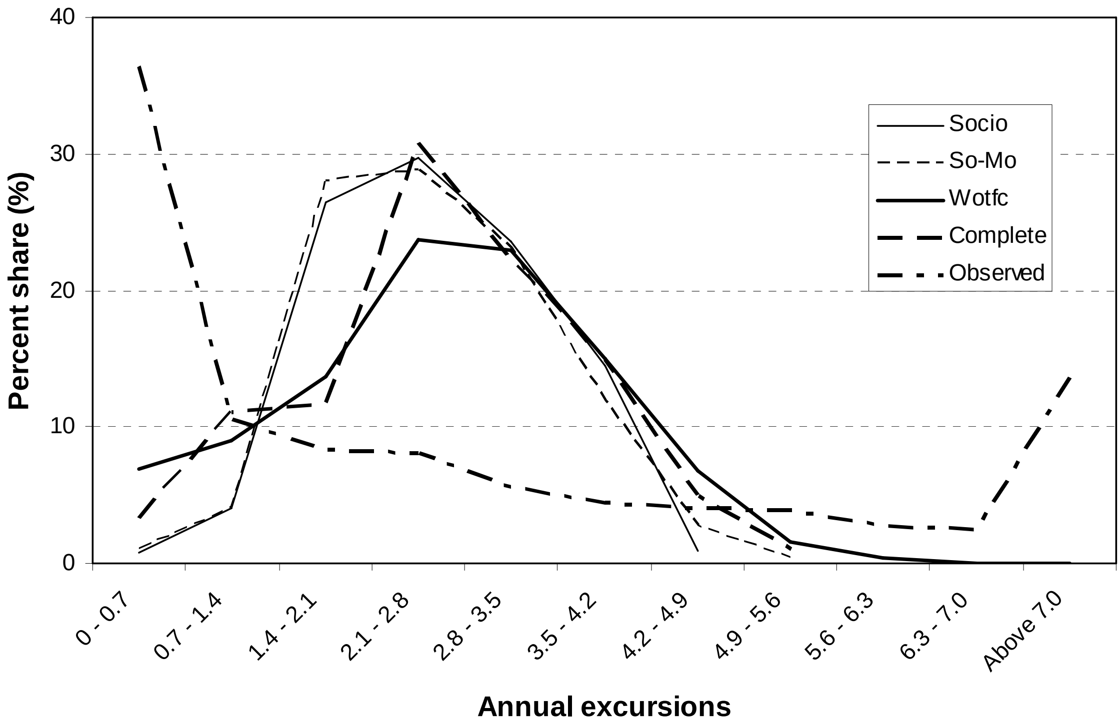


5.3.3 Travel market Switzerland 2001

The annual excursions distributions estimated using various regression models are shown in Figure 19. The "Wotfc" model based distribution has a lower peak and relatively higher extremes when compared to other distributions. Additionally in the Figure 19, the reported excursions projected to annual excursions with expansion factor range of 1 to 4 were also included. A comparison of observed excursions distribution with that of model based distributions has come up with the following inferences:

- Share of immobiles in the observed excursions are a way higher compared to the model based excursions. And so in case of share of persons with annual excursions above 7.
- The extremes of estimated excursions distribution differ slightly among the models.

Figure 19 Distributions of estimated average annual excursions (SRM 2001)



5.4 Excursions immobiles

One of the important outcomes that can be drawn from the results obtained is the impact on share of excursions immobiles reported to that of simulation-based projected excursions immobiles. Table 16 shows the detailed statistics on the share of excursions immobiles, both reported and projected by model type. From the tables it can be concluded that, the share of projected excursions immobiles has drastically dropped compared to that of reported excursions immobiles, which can be addressed to the generalisation of the excursions distribution using random coefficients. The two factors that significantly influence the reported share of excursions immobiles are reporting period and share of excursions in the reported travel information. But it is hard to draw some conclusions on either the individual or combined effect of these factors on the drop of excursions immobiles shares by model type. However, the generalisation of the excursions distributions has shown a significant change in the share of excursions immobiles.

Table 16 Summary on share of excursions immobiles[%] by model type

Description	EVE 1998	MZ 2000	SRM 2001
Reporting period (days)	14	1	365
Share of excursions [%]	17	02	43
Before	70	96.2	36.0
Socio	07	19.0	1.5
So-Mo	15	22.5	1.5
Wotfc	33	23.5	7.0
Complete	51	26.0	3.5

*: Less number of parameters were used than required

5.5 Model complexity and fit

Within the classical modelling framework, model comparison generally takes place by defining a measure of *fit*, typically a deviance statistic, and *complexity*, the number of free parameters in the model. Since increasing complexity is accompanied by a better fit, models are compared by trading off these two quantities. A model comparison using the Bayesian information criterion requires the specification of the number of parameters in each model (Kass and Raftery, 1995). A study was conducted on Bayesian measures of complexity and fit that can be combined to compare the models of arbitrary structure (Spiegelhalter *et al*., 2000). This study has come up with “Deviation Information Criterion (DIC)” to compare model complexity and fit. WinBUGS software was developed based on this study and the authors are the major contributors to the software development. Current study also uses the same concept implemented in WinBUGS, so that the output obtained from WinBUGS can directly be used in the analysis.

Some of the important statistical terms used in the model complexity and fit analysis are described as (Spiegelhalter *et al*., 2003):

Dbar: this is the posterior mean of the deviance, which is exactly the same as if the node 'deviance' had been monitored (WinBUGS generally automatically sets up a logical node to measure a quantity known as *deviance*). This deviance is defined as -2 * log(likelihood): 'likelihood' is defined as p(*y* | *theta*), where *y* comprises all stochastic nodes given values (i.e. data), and *theta* comprises the stochastic parents of *y* - 'stochastic parents' are the stochastic nodes upon which the distribution of y depends, when collapsing over all logical relationships.

Dhat: this is a point estimate of the deviance (–2 * log(likelihood)) obtained by substituting in the posterior means *theta.bar* of *theta*: thus Dhat = –2 * log(p(*y*|*theta.bar*)).

pD: this is 'the effective number of parameters', and is given by

pD = Dbar – Dhat.

Thus pD is the posterior mean of the deviance minus the deviance of the posterior means.

DIC: this is the 'Deviance Information Criterion', and is given by

DIC = Dbar + pD = Dhat + 2 ∗ pD.

The model with the smallest DIC is estimated to be the model that would best predict and replicate dataset of the same structure as that currently observed.

It is important to note that DIC assumes the posterior mean to be a good estimate of the stochastic parameters. If this is not so, say because of extreme skewness or even bimodality, then DIC may not be appropriate. There are also circumstances, such as with mixture models, in which WinBUGS will not permit the calculation of DIC.

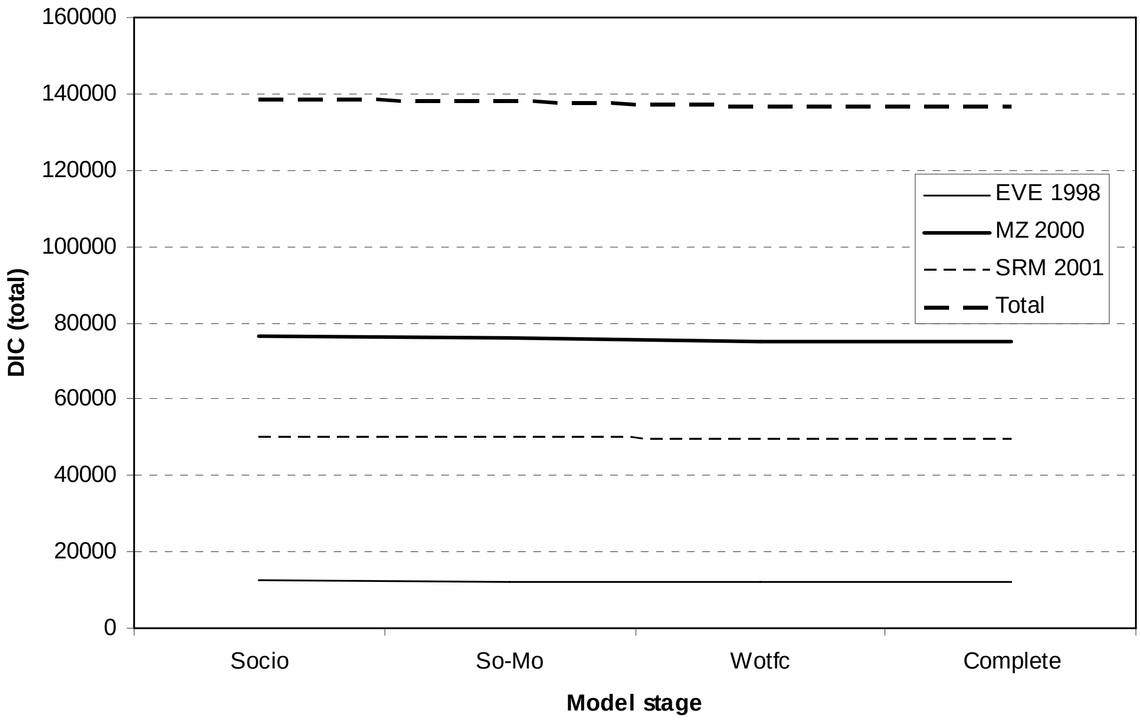
Table 17 shows the statistical summary of model fit analysis, a part of model output extracted from the WinBUGS. As mentioned before, the model with the smallest DIC is estimated to be the model that would best predict a replicate dataset of the same structure as that currently observed. Applying this to the current model, following conclusion can be made: even though the differences in total DIC among various models is small, with least total DIC, the “Complete” model fits better to estimate the excursion. Another interesting feature that the model fit analysis revealed is, model fits better with an addition of either one or a group of variables. Mathematically, model fit is proportional to the number of independent variables in the random-effect Bayesian regression model.

Table 17 Statistical summary of the model fit analysis

Model	Component	Dbar	Dhat	pD	DIC
Socio	EVE 1998	12273.7	12368.7	4.983	12378.6
	MZ 2000	76318.0	76312.3	5.711	76323.7
	SRM 2001	49766.6	49761.6	5.028	49771.7
	Total	138458.0	138443.0	15.723	138474.0
So-Mo	EVE 1998	12178.0	12171.0	6.981	12184.9
	MZ 2000	76002.8	75995.0	7.712	76010.5
	SRM 2001	49847.1	49840.1	7.050	49854.2
	Total	138027.0	138006.0	21.743	138050.0
Wotfc	EVE 1998	12041.1	12030.0	11.055	12052.1
	MZ 2000	74895.0	74866.0	29.007	74924.0
	SRM 2001	49577.3	49552.3	25.046	49602.4
	Total	136513.0	136448.0	65.108	136579.0
Complete	EVE 1998	12036.8	12027.0	9.783	12046.6
	MZ 2000	74864.3	74849.0	15.344	74879.7
	SRM 2001	49572.3	49557.7	14.563	49586.9
	Total	136473.0	136434.0	39.690	136513.0

Figure 20 shows the DIC distributions at various stages of the model.

Figure 20 Distributions of total DICs against different stages of the model



Model complexity is generally measured with the number of free parameters. Free parameters are the additional parameters that were defined in the model.

Free parameters = (Number of parameters defined in the model – Number of effective parameters)

Model complexity is all about defining the optimum number of parameters required to fit the model for the real-time problems. A model with less number of free parameters is less complex compared to the model with many free parameters. Table 18 shows the statistical summary of the model complexity analysis. WinBUGS generate the effective parameters necessary to fit the corresponding model to estimate the excursions. In the model complexity analysis, number of free parameters was calculated and the models were ranked according to their complexity.

Table 18 Statistical summary of model complexity analysis

Model	Total parameters used in the model	Effective parameters	Free parameters	Complexity rank
Socio	24	16	8	1
So-Mo	36	22	14	3
Wotfc	46	65	-19*	4
Complete	51	40	9	2

*: Less number of parameters were used than required

Following conclusions can be drawn from the model complexity analysis:

- Socio model with least number of free parameters has emerged as the model with least complexity
- Adequate number of parameters was not specified in the "Wotfc" model. A parameter redundancy was noticed.
- Complexity has nothing to do with the number of model parameters. In other words, increase in number of model parameters need not necessarily increase the complexity. The "Complete" model case is an example.

6. CONCLUSIONS

Three travel surveys each from corresponding travel surveys series were selected to test the combining the travel information observed from multiple sources. A common thread of travel information was aimed at. The preliminary data mining has resulted the common travel information in the form of “Excursions”: a home-based round trip that lasts a minimum of 4 hours and a maximum of 24 hours with distance to the destination is at least 25 km.

Statistical models in the form of linear regression models were chosen to combine the travel information. To account the unobserved effects outside the model variables, it was decided to introduce random coefficients to the linear regression models. The random-effect linear regression models development process in the WinBUGS environment involves construction of graphical models using DoodleBUGS editor and validating the model for syntax. The syntactically correct models were updated with MCMC simulation using Gibbs sampling for a million times each. All the models were converged at 10000 updates for two chains. Three forms of output extracted from the updated model are:

- Specifications of the average posterior distribution of the fixed and random effect parameters.
- Pair-wise correlations among the parameters
- Deviation information criteria (DIC) for model complexity and fit analysis

A correlation matrix was constructed using the pair-wise correlations. With the average posterior distributions specifications and the corresponding correlation matrix, correlated synthetic data was generated for all the parameters. Using the specified regression models, for the observed travel associated data and posterior distributions based synthetic parameters data, new excursions distributions were calculated for both individual data sources and the combined. Comparison of the reported excursions distributions and the new posterior distributions based derived excursions distributions revealed some interesting insights such as:

- The new derived excursions distributions fit better than that of the extrapolated reported excursions distributions.
- The combined excursions distributions based on the joint posterior distributions differ from the derived excursions distribution of any individual data source and had the influence from all the data sources.
- The survey characteristics have severely affected the derived excursions distributions and smoothened the combined excursions distributions.
- The derived and combined distributions of the models “Wotfc” and “Complete” are much more realistic.

To understand the effectiveness among the models, model complexity and fit analysis was performed. Following conclusions can be drawn from the model complexity and fit analysis:

- Both the models “Wotfc” and “Complete” fit better than the rest of the models.

- The “Socio” model has the least complexity
- The “Complete” model fits better when compared to the other models.
- Considering both the model complexity and fit, “Compete” model was found the best model.
- The number of parameters and variables has no impact the model complexity and fit analysis.

The final conclusion is on the immobiles analysis. Extrapolation of the immobiles share for longer durations has severe impacts on the overall distributions. Updating the model has certainly brought down the immobiles share to that of reported immobiles share. One of the biggest advantages of combining the travel information is a successful generalisation of the diverse reported excursions distributions obtained from disparate sources.

7. REFERENCES

Chalasani, V. S. and K.W. Axhausen (2003) A comparison of Swiss household travel surveys: EVE1998, MZ2000, and SRM2001, *Arbeitsbericht Verkehrs- und Raumplanung*, **174**, Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich, Zürich.

Bieger, T. and C. Lässer (2002) *Travel Market Switzerland* –Basic Report and Variables Overview, Institute for Public Services and Tourism, St. Gallen.

WinBUGS 1.4.1 (2005) http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

Gilks, W. R, S. Richardson, and D.J. Spiegelhalter (Eds.) (1996) *Markov Chain Monte Carlo Methods in Practice*

Kass, R. and A: Raftery (1995) Bayes factors and model uncertainty. *Journal of the American Statistical Association*, **90**, 773–795.

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.

Murphy, K. (2004) A Brief Introduction to Graphical Models and Bayesian Networks. MIT webserver: http://www.ai.mit.edu/~murphyk/Bayes/bnintro.html#repr

Richardson, A.J., E.S. Ampt, and A.H. Meyburg (1995) *Survey methods for transport planning*. Eucalyptus press.

Shafer, J.L. (2000) *Analysis of incomplete multivariate data*. Monographs on Statistics and applied probability, no. 72. Chapman & Hall.

Simma, A. (2005) Methodik und Hauptergebnisse, *Reiseverhalten der Schweizerische Wohnbevölkerung 2003*, BFS; Neuchâtel.

Spiegelhalter, D. J., A. Thomas, and N.G. Best (2003). *WinBUGS Version 1.4.1 User Manual*. MRC Biostatistics Unit

Spiegelhalter, D. J., A. Thomas, and N.G. Best (1996). Computation on Bayesian graphical models. *Bayesian Statistics* 5 pp. 407—425

Spiegelhalter, D.J, N.G. Best, B.P. Carlin and A. van der Linde (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society* B. 64, 583-640.

Swiss Federal Office for Spatial Development and Swiss Federal Statistical Office (2001) Travel behaviour results from the 2000 travel Microcensus, *Mobility in Switzerland,* BFS and ARE, Bern.

Swiss Federal Statistical Office (2002), Methodologies and principle results, *Travel Behaviour of People Living in Switzerland in 1998*, BFS, Neuchâtel.