

OPUS

Optimising the use of Partial information in Urban and regional Systems

Project IST-2001-32471

ITS Programme

TRANSPORT

Title : **Feasibility Studies : Belgium and Regione Lombardia**

Author(s) : Cornelis, E. (FUNDP-GRT)
Ponti, C. (Systematica)
Logie, M. (Imperial College London, Minnerva)

Deliverable No. : D10.3
Version : 2.1

Contract Date :
Submission Date :

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

Project Coordinator : Imperial College London
Contact Person : John Polak
Address : Imperial College
South Kensington campus
London SW7 2AZ
United Kingdom
Telephone : +44-(0)20-7594.6089
Fax : +44-(0)20-7594.6102
e-mail : j.polak@imperial.ac.uk

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

TABLE OF CONTENTS

<i>Optimising the use of Partial information</i>	1
<i>in Urban and regional Systems</i>	1
Executive Summary	4
1. Introduction and Framework	5
1.1 The Feasibility Studies.....	5
1.2 Objectives of the Feasibility Studies.....	5
1.3 Objective of the Deliverable	6
1.4 Structure of the Deliverable	6
2. Feasibility Study for Regione Lombardia	7
2.1 Disaggregation Problem Statement.....	7
2.2 Definition of the area	7
2.3 Steps in the OPUS Methodology	9
2.4 The GAPM.....	9
2.5 Sources of Data.....	10
2.5.1 The “Territory” area	11
2.5.2 The “Transport Infrastructure” area	11
2.5.3 The “Traveller behaviour” area	12
2.6 The BBN.....	12
2.7 The Multi-dimensional Data Cube	14
2.7.1 Specification	14
2.7.2 Application of Dominici Method	17
2.7.3 Scalability	17
2.8 Traffic Flow Data.....	18
3. Feasibility Study for Belgium	19
3.1 Problem Statement.....	19
3.2 Previously used deterministic approach	19
3.3 Applying the OPUS Methodology	23
3.4 The GAPM.....	23
3.5 Sources of Data.....	24
3.5.1 For the population:.....	25

3.5.2 For the mobility behaviours:.....	25
3.6 The BBN.....	25
3.7 Application of Dominici Method	28
4. Summary and Conclusions.....	30
4.1 Regione Lombardia Feasibility Study.....	30
4.2 Belgian Feasibility Study	30

EXECUTIVE SUMMARY

This deliverable D10.3 is an element of Work Package WP10 of the OPUS project. Work Package WP10 has the title: “Feasibility Studies”.

The aims of this work package are to:

1. Examine issues that exist for transport planning data synthesis in other European cities and in a national context
2. Provide inventories of data sources and gaps in data
3. Review potential operation of the methodologies with local officials
4. Design an approach in each feasibility study area and report

This deliverable D10.3 provides the final report of the Work Package and describes the potential application of the methods developed and researched by OPUS to the Feasibility Studies in Belgium and in Italy, located in Regione Lombardia. This report builds on a first report [D10.1] that presented a specification of the feasibility studies and [D10.2] reported on available the data sources and gaps, which is a necessary starting point for the application of OPUS.

The Feasibility studies have shown how the data may be cast into a form that allows the OPUS methods (see report [D4.2 Supplement] and software deliverables [D7.2]) to be applied and whose operation has been demonstrated by other work packages (reports [D8.2] and [D9.3]).

The feasibility of the procedure has been confirmed, but the need for caution in the problem size (notably in respect of the total number of data dimensions) and the desirability of using OPUS methods in association with focused local surveys, together with the larger sets of surveys that have been considered by the Feasibility studies.

1. INTRODUCTION AND FRAMEWORK

1.1 The Feasibility Studies

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 OPUS partner is participating in the technical work of OPUS, contributing to methodology development, commenting on reports from others and participating in the User Forum and technical meetings. At the appropriate point in the programme, selected local partners are to undertake a substantial review of the feasibility of applying the methodology in their area.

The participants in the OPUS project, identified to conduct feasibility studies in the domain of transport are:

- FUNDP, Transport Research Group (University of Namur), Belgium
- Systematica, Italy.

1.2 Objectives of the Feasibility Studies

The feasibility studies will evaluate potential application of the OPUS Methods. The available data, institutional processes, philosophy and methodological approaches to transport planning are very different between parts of Europe. The feasibility studies will identify and document the issues, explore the data sources in detail and identify other data needed for a successful implementation. If the OPUS methodology requires adaptation to meet local conditions, this too can be specified.

At present, it is anticipated that the feasibility studies will conclude with designs for implementation projects to generate a transport information synthetic database and updating process. However, it is not precluded that some authorities may see the benefits and wish to conduct the projects. This could be done independently or alongside an extension of the OPUS work.

The project is undertaking pilot and feasibility study transport 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 that will be available in the proposed potential applications are:

- 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;

- Avoidance of confusion from different, apparently conflicting, estimates of the same quantity

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 feasibility studies are to examine how the OPUS methodology might apply in particular local settings.

As described in the OPUS Deliverable Report D10.2, there exists a considerable range of transport-related data for the Regione Lombardia, including the important area of Milan. Part of the data is the major origin destination survey that was undertaken in 2002 for the region. The OPUS methodology is of value for its ability both to exploit the diverse set of data and also to impute information that allows data to be synthesised at a spatially more detailed level. Typical standard approaches for the disaggregation of trip matrices are based on iterative proportional fitting methods that require seeding of the detailed matrices for which data is to be imputed. This seeding produces a result but is often arbitrary in nature. The OPUS approach makes full use of data to calculate the parameters of distributions that relate to the expected values, and which takes full account of uncertainty.

For the Belgian feasibility study, the purpose is the determination of mobility profiles for the whole Belgian country. The first step is building a synthetic population for Belgium, at municipalities level (i.e. 589 entities). Then, next step is associating mobility profiles with these populations. This process is further completed with forecasts based on demography evolution models. As explained in the OPUS Deliverable Report D10.2, the used data sets present some incoherencies and therefore the results provided by the applied deterministic method are also limited due to this drawback. The OPUS methodology is a good opportunity to investigate how it is possible to deal with these uncertainties, to jointly exploit these different data sources and to use different margin sums in order to fill in the gaps in cells (this means imputing information for disaggregated categories of population at spatially detailed level). One of the main issues will be the scale of the problem as its has already been a serious drawback when implementing the deterministic methodology.

1.3 Objective of the Deliverable

This deliverable document provides a description of how the OPUS methodology, as defined in the OPUS documents D4.2 and D4.2 Supplement (and related documents), can be applied to the problem statement for the transport feasibility studies given in D10.1 and D10.2, which also included an assessment of the available data and information sources.

1.4 Structure of the Deliverable

This deliverable is twofold: first, the Regione Lombardia feasibility study is described, then the Belgian feasibility study is presented. In each case, we first state the problem and then described how this case could be expressed using the OPUS methodology, as defined in the OPUS documents D4.2 and D4.2 Supplement (and related documents). The same steps are considered for both feasibility studies: description with a General A Priori Model (GAPM) which is then more formally translated in a Bayesian Belief network (BBN). We also indicate, in both situations, how the Dominici method could be useful for estimation of unknown disaggregated quantities, knowing more aggregated sums.

2. FEASIBILITY STUDY FOR REGIONE LOMBARDIA

2.1 Disaggregation Problem Statement

The objective for the Lombardy Region Feasibility study, as defined in document D10.2, is to focus on the topic of disaggregating transport trip matrices, that is, **how to obtain small area travel information from a regional area information database and to consider how the OPUS methodology can combine available data sources to obtain such a disaggregation.** The Feasibility study has investigated the data required to perform the disaggregation by journey purpose and travel mode.

2.2 Definition of the area

While the context of the Feasibility study is provided by information collected across the large area of the Regione Lombardia, the study area for the Feasibility study has been focused at the boundary of the Municipality of Milan, on the east side of the city, as shown in Figure 2.1.

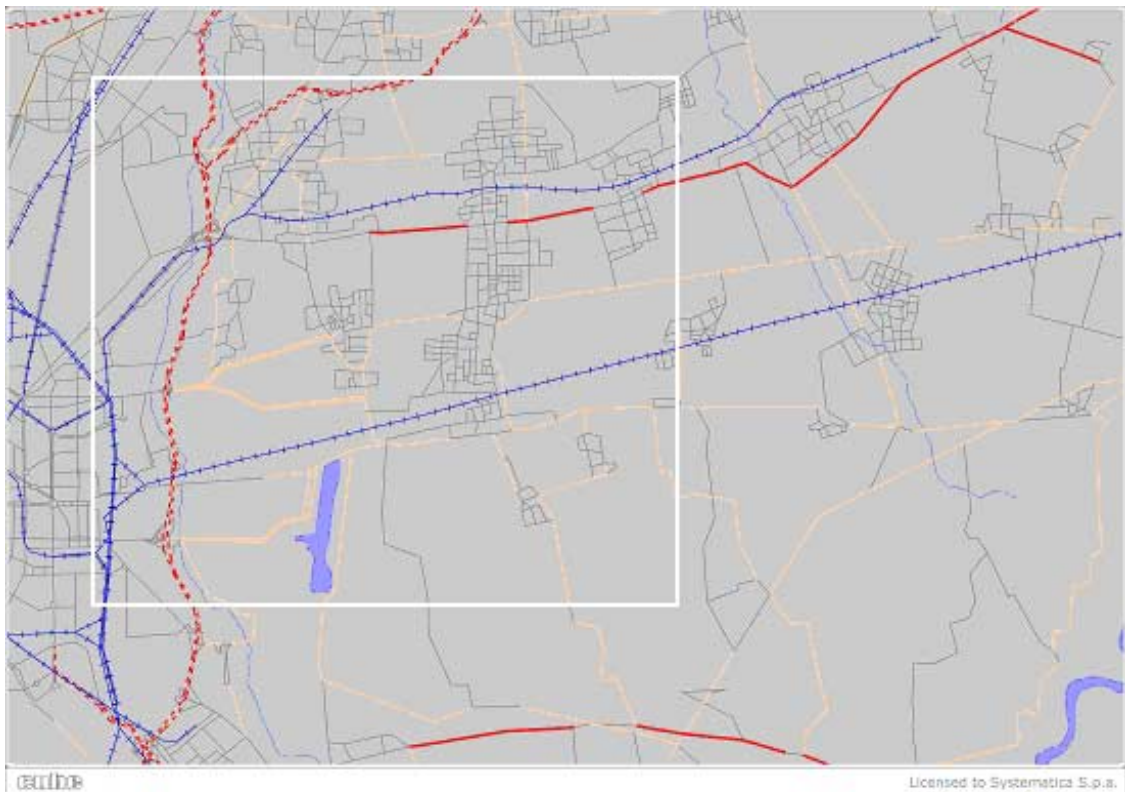


Figure 2.1 Study Area for Disaggregation Problem

The decision to choose the area delineated by the white line in Figure 2.1 is related to following reasons:

- The area considers not only the city of Milan but also the boundaries of the municipality and the Provincia; this offers more opportunities to involve different public administration levels;
- In this area some significant road improvement project will be implemented in the following 10 years; this kind of scenarios challenge not just the city administration but also the Provincia and Regione;

- A new suburban train stop has been recently activated. This line will connect Varese with Treviglio (two different provinces that connect with provincial di Milano). The nature of the suburban train connections in and around Milano is shown in Figure 2.2.

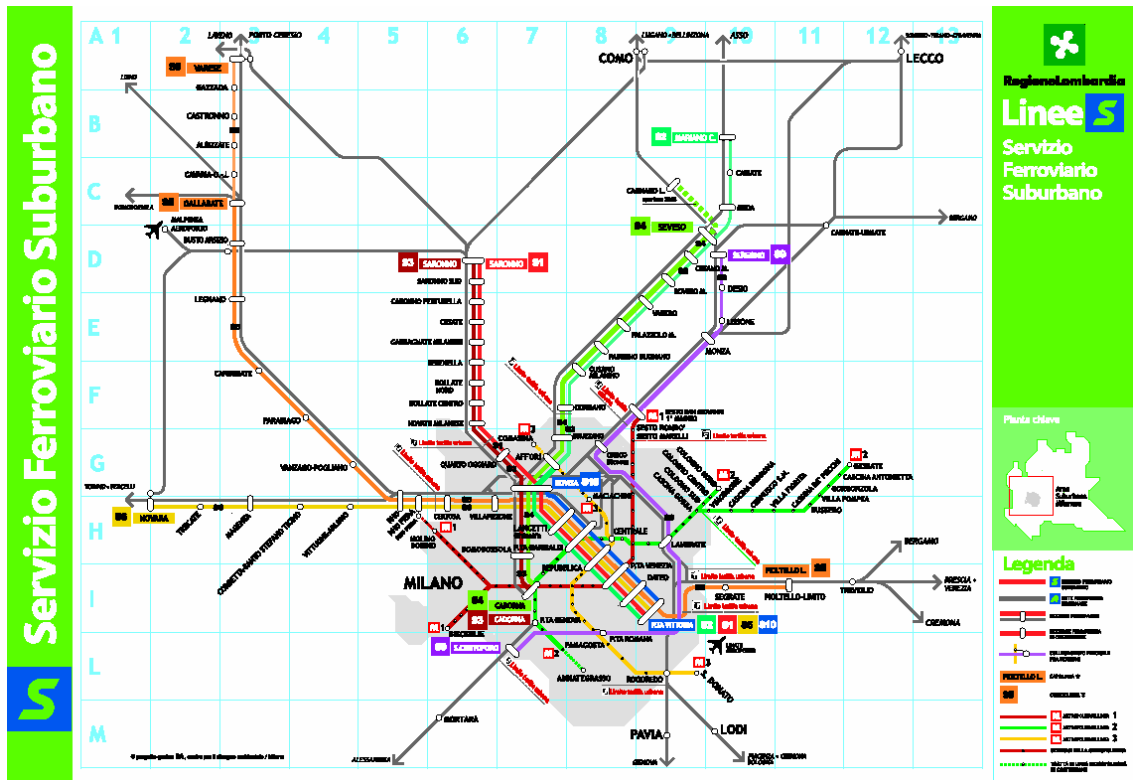


Figure 2.2 Suburban Rail Connections

- In the same area a new general office building master-plan is projected;
- For the purposes of understanding the application of the OPUS methodology, it is easiest to consider the study area defined above in Figure 2.1 in the more schematic form shown below in Figure 2.3.

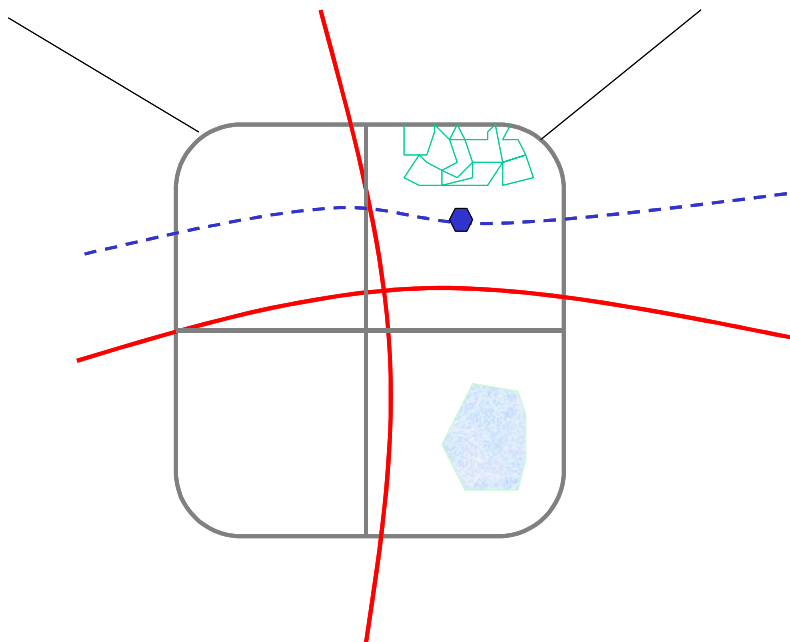


Figure 2.3 Schematic View of Disaggregation Study Area

2.3 Steps in the OPUS Methodology

The OPUS methodology is applied through a set of steps that relate the problem specification and available information to a set of conceptual abstractions that generate a holistic view of the system being considered that can then be converted to the more formal and precise formulation of a Bayesian Belief Network (BBN). However, the interest that the Lombardia Regione Feasibility study has in producing detailed, disaggregated estimates from the broader regional datasets for the much smaller district of Milan identified in Figure 2.1 means that the formation of the data into a contingency table, corresponding to a multi-dimensional cube, is an important feature of this Feasibility study.

2.4 The GAPM

A starting point for the OPUS methodology is the derivation of a GAPM (Generalised A Priori Model). This is a graphical expression for the domain of interest – here transport – that identifies the principal components of interest and their interactions. That means mainly the data and the modelling sources.

The definition of the GAPM and its precision of detail provide the basis to organise the concepts, experience and information that can guide the construction of a Bayesian Belief Network.

The data analysis and consideration of the disaggregation problem that is the focus of the feasibility item in the Lombardy Region has led to a GAPM which considers three macro-elements which describe the interaction between the Region's territory and transport to the extent required, as shown in Figure 2.4:

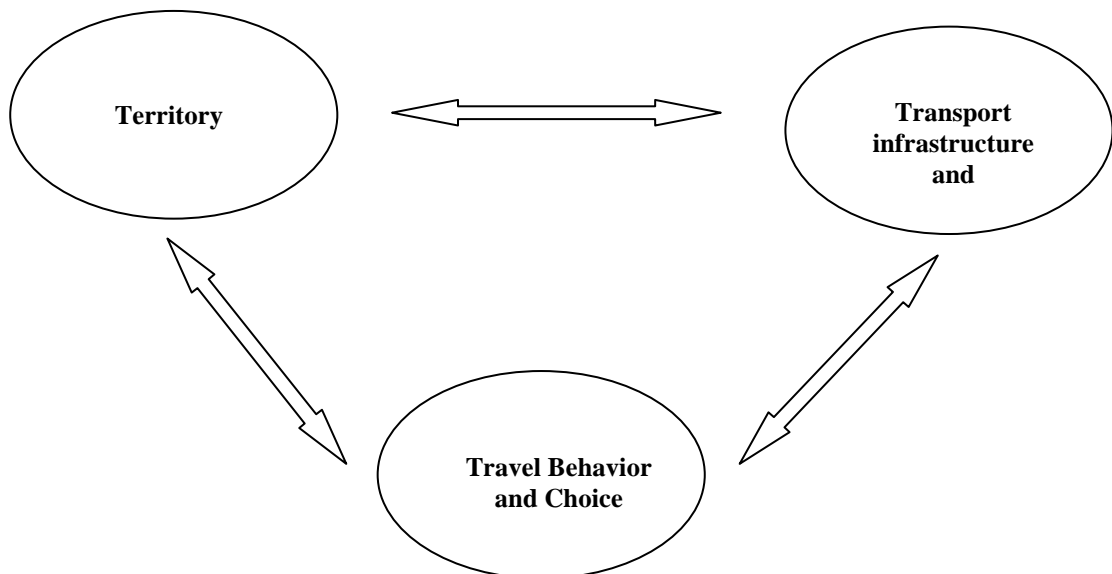


Figure 2.4 Primary Components for Lombardy GAPM

Territory: this covers all the characteristics of the territory such as Land-use, Population and relative socio-economic characteristics. For simplicity, time is not considered explicitly and so the data implicitly refers to the all-day situation for a typical day and season. The GAPM thus considers all the potential population including inhabitants (as Home work person trip in an AM peak hour model), or employees on a trip for work purpose (work to work for a meeting), or can be students who reaches a leisure park from school). It is important to distinguish POPULATION as inhabitants given by the CENSUS DATABASE with socio-economic characteristics, and other potential populations of people who can form an attraction or a generation TRIP-END for different time periods.

Transport Infrastructure and performances: this is comprehensive over all the different transport choices, including data on their operational performance (varying by time of the day, etc.). The information on the infrastructure is suitably considered as being provided by an (existing) multi-modal transport model with interchange functionality between the highway system (LGV, car, motorbike, etc.), public transport (train, buses, taxi), and pedestrian trips and bicycles. In this context such network models effectively represent transport infrastructure inventory databases.

The **Travel Behaviour and choice** box then contains the relationships between the other two areas that enable the description of the interactions between territory and transport effects.

The broad view of the GAPM presented in Figure 2.4 can be elaborated into a more detailed GAPM, as shown in Figure 2.5, which also indicates the sources of data for information on different aspects of the GAPM for the Lombardy Region. The sources of data are described further in the following section.

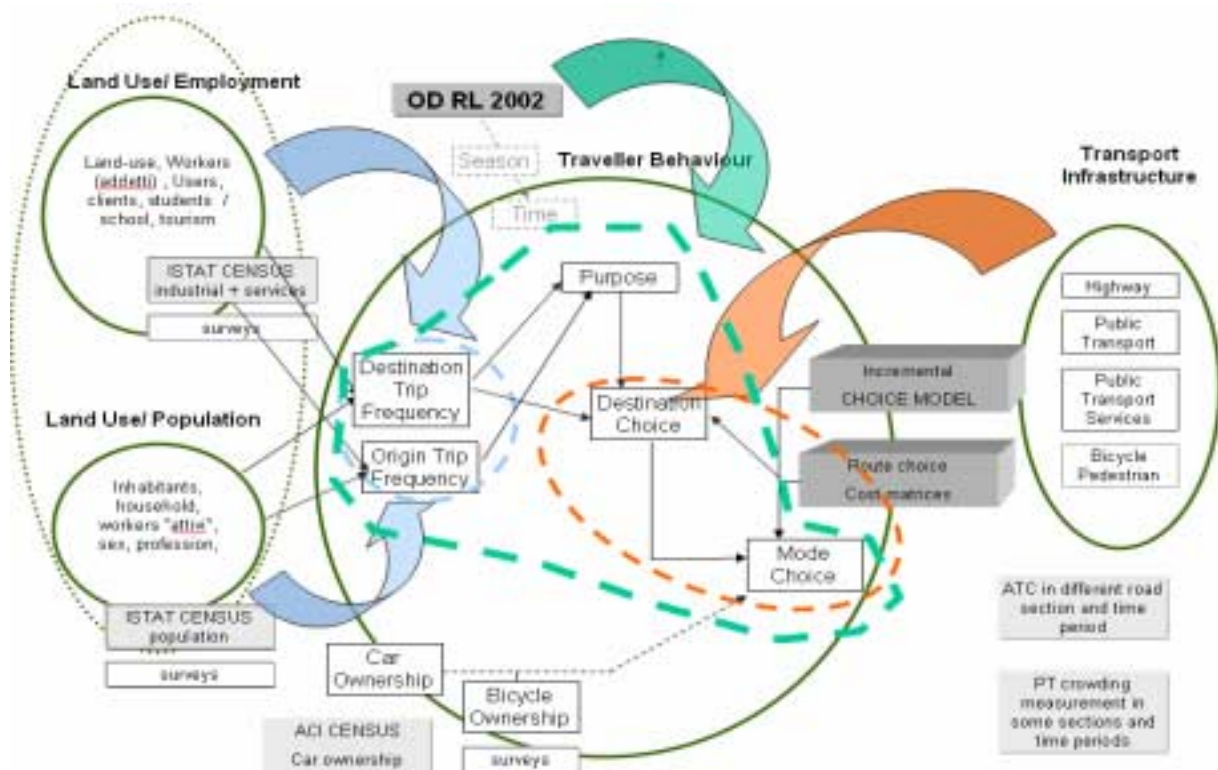


Figure 2.5 GAPM for Disaggregation Problem

2.5 Sources of Data

With reference to the views of the topic domains identified by the Lombardy Region Feasibility study given in Figure 2.4, we now consider the information sources that can be used to generate the conditional probability distributions required by the Bayesian Belief Network description.

2.5.1 The “Territory” area

The territory can be described by the land use in a qualitative and in a quantitative way. The following categories can be defined that consider land use in terms of the activities that generate and attract trips and which affect trip frequency:

- Home
- School
- Activities:
 - Intensive and fixed job (many workers with fixed time rule), no significant extra trips visiting the location (Fiat);
 - Intensive and fixed job (many workers with fixed time rule), with significant extra trips visiting the location (i.e.: Hospital)
 - Non Intensive and fixed job (workers with fixed time rule), with significant extra trips visiting the location (i.e.: bank offices and other services)
 - Non Intensive and not fixed job (workers without fixed time and location rules), with significant extra trips visiting the location (e.g. IBM)
- Leisure

The **Home Land-use** implies we have a Population dataset which will contains the number of people, by age, by sex, by professional capacity, possibly by income, by car-ownership and bicycle-ownership. The **School land-use** implies knowledge of how many workers can be at school, by time of the day, and by day, and how many students by age and sex are supposed to be present at a certain time and day. In a similar way also work activities and leisure activities can be described as potential traveller quantities (Trip-ends).

The Italian Census data (Censimento della POPOLAZIONE) describes the Home land-use in detail for all the people and their socio-economic characteristics. In a similar way there is a census data (Censimento delle INDUSTRIE) which refers to fixed workers in their fixed work location.

Information about quantities and types of trip ends for other activities, school, and leisure is not available and would need to be constructed either with a direct survey or estimated through trip frequency modelling.

The Census data is available at a fine level of zoning, which provides a convenient unit for the defining the spatial detail required by the disaggregation problem.

2.5.2 The “Transport Infrastructure” area

In order to construct and verify relationships between territory and transport it is necessary to describe as well as possible the transport system. In consideration of the different mode choices it will be necessary to know and describe a highway network and its performance, the public transport network and the offered services (and their performance), as well as other modes like pedestrian and bicycle.

Information is also required on route choices and the consequential flows on links (by mode) and related travel costs.

Information about the transport infrastructure itself corresponds to that of inventory databases, but information on usage of the infrastructure, and by whom and for what purpose requires different dataset can be used to give a partial knowledge of part of this system:

- Direct survey (by interviewing travellers);
- Automatic traffic count

- Season and electronic ticketing, ticket with identification of on – off stop
- Numbers of travellers on a bus/train in a section
- Etc..

These datasets are frequently either spatially or temporally restricted, and are intended to give good partial information about some modes and a sub-area of the system.

2.5.3 The “Traveller behaviour” area

Considering travel by persons, each trip is related to a person: not all the persons are supposed to be travellers, a traveller can make one ore more trip. A trip has following characteristic:

- An origin,
- A purpose,
- A destination
- A frequency
- A time / season
- A Mode choice.

The travel can be thought of to consist of trips made by an individual i from a population with certain socio-economic characteristics s , each with an origin $Orig$, a destination $Dest$, a travel purpose $Purp$, a travel mode $Mode$, a path $Path$, a cost for that path LOS .

We can think of the *population* of interest as a list $L_0 = \{i, Orig, s\}$ of individuals, each with a unique number i , an origin zone $Orig$, and socio-economic characteristics s .

The individuals of this population may decide to embark on a trip; those that do enter the list $L_1 = \{i, Orig, Dest, S, Path, LOS, Purp, Mode\}$ of *person-trips*. Person-trips are made by individual i who travels at a certain time, on a certain date for a certain *purpose* from an *origin*, to a *destination*, along a *path* which has a certain travel *mode* and offers a certain *level of service*.

Regarding the Lombardia feasibility study, the 2002 OD survey, done by the Regione (‘OD RL 2002’), gives a generally comprehensive and precise description of the relation between people and transport. Although this regional dataset has very many records, the sampling is only sufficient to make it statistically reasonable to associate this data with relative large traffic zones.

2.6 The BBN

In seeking to transform the GAPM (see Figure 2.5) into a BBN we need to define exactly which (mathematical) relationship defines all the elements of the transport system.

The BBN is formed by following the flow of interactions expressed in the GAPM (Figure 2.5) and forming it into a directed graph. This graph may have branches, but the specific objective of this feasibility means that the graph has a largely linear form.

At the top we have n_{is} , the number of people living in zone i with individual characteristic s . These people decide how many trips to make, and for what purpose. This is modelled by calculating the probability $P_{pk|is}$ of making k trips for trip purpose p given that they are in zone i and have user characteristics s .

We can assume that the probability $P_{pk|is} = \mathbf{f}(s, p, \text{time, accessibility})$

The number of people making a trip for purpose p from origin zone i and user characteristics s is then calculated as:

$$n_{ips} = n_{is} \sum_k P_{pk|is}$$

Next step suppose to define the distribution of the trips in different destinations j , depending from purpose and accessibility. So:

$$T_{ijps} = n_{ips} P_{j|ips} \quad \text{where } P_{j|ips} = \mathbf{f}(s, p, \text{time, accessibility})$$

Next the probability $P_{m|jips}$ is calculated of choosing and transport mode m given that one is embarking on a trip for trip purpose p from origin zone i to destination j . Then, referring to a O-D matrix we can consider:

$$T_{ijmps} = T_{ijps} P_{m|jips} \quad \text{where } P_{m|jips} = \mathbf{f}(s, p, \text{cost matrices})$$

Based on the O-D flows and the route choice probabilities between O-D pairs $P_{\pi|ijmps}$, we can calculate the route-flows as:

$$f_{ij\pi mps} = T_{ijmps} P_{\pi|ijmps}$$

With the network structure coded as a link-path incidence matrix \mathbf{A} , this gives us a relationship between path flows and link flows:

$$v_{\pi mps} = \mathbf{A} \circ f_{ij\pi mps}$$

The link flows $v_{\pi mps}$ determine the traffic speed on the links, which leads to link travel times:

τ_{am}

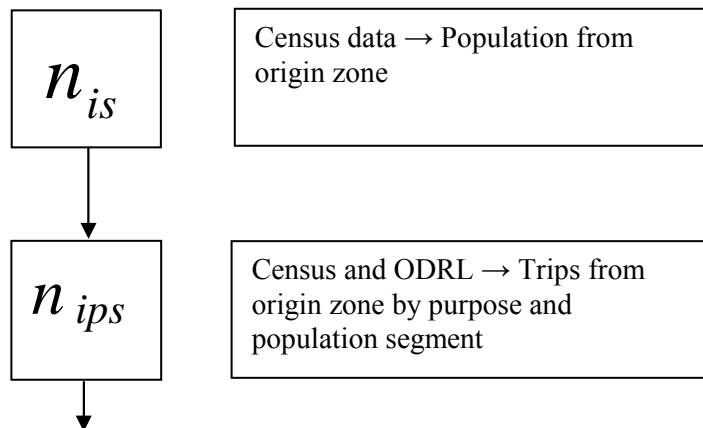
$$\tau_{am} = g(v_{am})$$

$$v_{am} = \sum_{ps} v_{amps}$$

Summing the link travel times across paths gives path travel times $\tau_{ij\pi m}$

$$\tau_{ij\pi m} = \sum_{a \in \pi_{ij}} \tau_{\pi m}$$

The path travel times $\tau_{ij\pi m}$ contribute to the total (generalised) path costs $C_{ij\pi m}$, which in turn determine the probability of a certain path π being chosen.



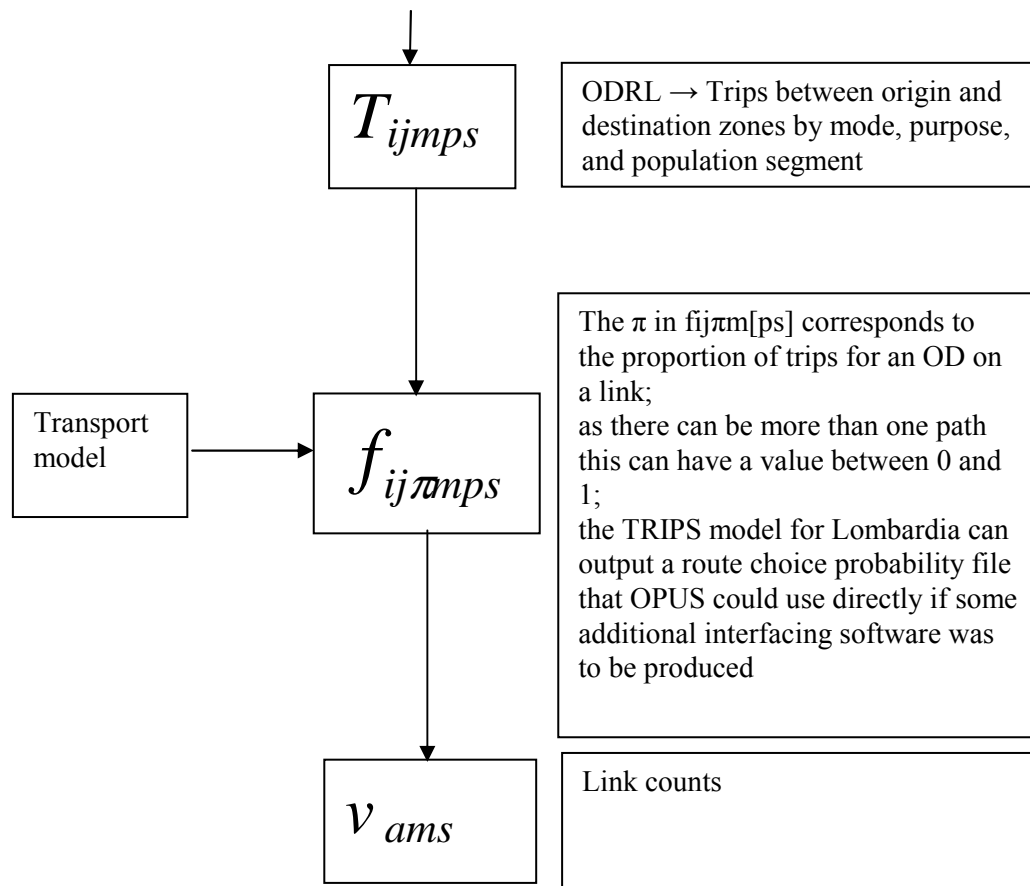


Figure 2.6 Bayesian Belief Network (BBN) for Disaggregation Problem

2.7 The Multi-dimensional Data Cube

2.7.1 Specification

The data associated with the central boxes in Figure 2.6 may be considered as a multi-dimensional data cube (or contingency table) whose dimensions are indicated by the subscripts. In the case of n_{ips} and T_{ijmps} the principal issue of concern to the Regione Lombardia Feasibility study is that information is not known from the ODRL survey for n_{ips} and T_{ijmps} at the level of spatial detail that is required.

This may be interpreted as meaning that data for certain (spatial) dimensions is missing. It is therefore of interest to consider the feasibility of applying the Dominici method [D4.2 Supplement] for addressing this problem.

Further to the discussion in Section 2.5.3 above, the different dimensions that relate to the Regione Lombardia Feasibility study can be identified as the following:

- Zone system (census / traffic) *
- Socio-economics characteristics (population segments)
- Purpose
- Time of day / season *
- Frequency

- Mode choice:
 - main mode *
 - single mode /by segment (HW, PT, bike or pedestrian)
 - Multimodal trip (combination of modes)

As noted, the important element in the multi-dimensional data cube for this application is the disaggregation of the spatial dimension, so we require two versions of these corresponding to large (aggregate – Z) and small (disaggregate – z) zones, and also two versions corresponding to home location (i, and I). The data lists would then become of the form :

$$L = \{i, I, \text{Orig}_z, \text{Orig}_Z, \text{Dest}_z, \text{Dest}_Z, S, \text{Path}, \text{LOS}, \text{Purp}, \text{Mode}\}.$$

For OD RL 2002 information, the list is

$$L = \{+, I+, \text{Orig}_Z, +, \text{Dest}_Z, S, +, +, \text{Purp}, \text{Mode}\},$$

From ISTAT Census information data on population and employment respectively provides lists of the form:

$$L = \{i, I, \text{Orig}_z, \text{Orig}_Z, +, +, S, +, +, +, +\} \quad \text{and}$$

$$L = \{i, I+, +, \text{Dest}_z, \text{Dest}_Z, S, +, +, +, +\}.$$

For the case of work and education trips these lists have more information of the form

$$L = \{i, I, \text{Orig}_z, \text{Orig}_Z, +, +, S, +, +, \text{Purp}=\text{Work|Education}, +\} \quad \text{and}$$

$$L = \{i, I+, +, \text{Dest}_z, \text{Dest}_Z, S, +, +, \text{Purp}=\text{Work|Education}, +\}.$$

Information about person type (S) that is common to the ODRL and Census data sources is provided by several person type attributes. The availability of the ‘Professional Category’ (relating to employment type and education level) is of particular interest to this application and provides a key means of linking these two data sets, but person age and gender attributes may be considered as well.

It must be noted that the Census deals in units of persons, whereas the ODRL considers person trips for a given time period. It is necessary to use a factor of trip rates by time period to ensure consistent units between these data sets. This information has been shown to be relatively consistent over time and can be derived from the ODRL.

The precision of the linking would be improved if information were available that related person type, purpose and mode at the detailed level, that is, data in the form

$$L = \{i, +, +, +, +, +, S, +, +, \text{Purp}, \text{Mode}\}.$$

This data is not available but we can observe that this might reasonably be estimated (modelled) from a general model applied to detailed land use information for the area of interest in Milan to give the numbers of people of type S, living in zone i and going shopping by mode car (for the given time period), and so on.

It is convenient to consider the trip data presented above in list form as being organised in the form of a multi-dimensional data cube. Not least for purposes of visualisation, the cube is shown as having three dimensions, though the number of data dimensions is not restricted in this way and, indeed, it is suitable to consider higher numbers of dimensions in combinations that result in three (combined) dimensions.

The data cube with the combined dimensions, and the correspondences with the list above is shown in Figure 2.7 below and we now describe the various elements of this diagram.

Firstly, it should be noted that the diagram relates to a single ODRL zone. This is denoted as (I, Z), with the interpretation varying according to whether the context implies home zone, origin zone, or destination zone. For clarity of presentation, we confine ourselves to

describing an origin or home zone, but the ‘OD Combination’ dimension can include destination zones, as well as origin-destination zone pairs.

The computation would be done over all ODRL zones in turn that cover the region of interest in Milan (as discussed in Section 2.2 above).

It is assumed, for simplicity, that each ODRL zone corresponds to a set of Census zones (equivalently denoted as (i,z)) – the possibility of ODRL zone boundaries not being contiguous with Census zone boundaries can be admitted but this would require additional factors to be incorporated.

The travel attributes are considered, as in the list above, to be Mode and Purpose (Purp), although this may be extended to include any further attributes offered by the travel survey.

The main (rectangular) cube in Figure 2.7 provides the target disaggregate set of information, and the ‘slabs’ shown beneath and at one end correspond to the totals across dimensions that are provided by the indicated sets of survey information.

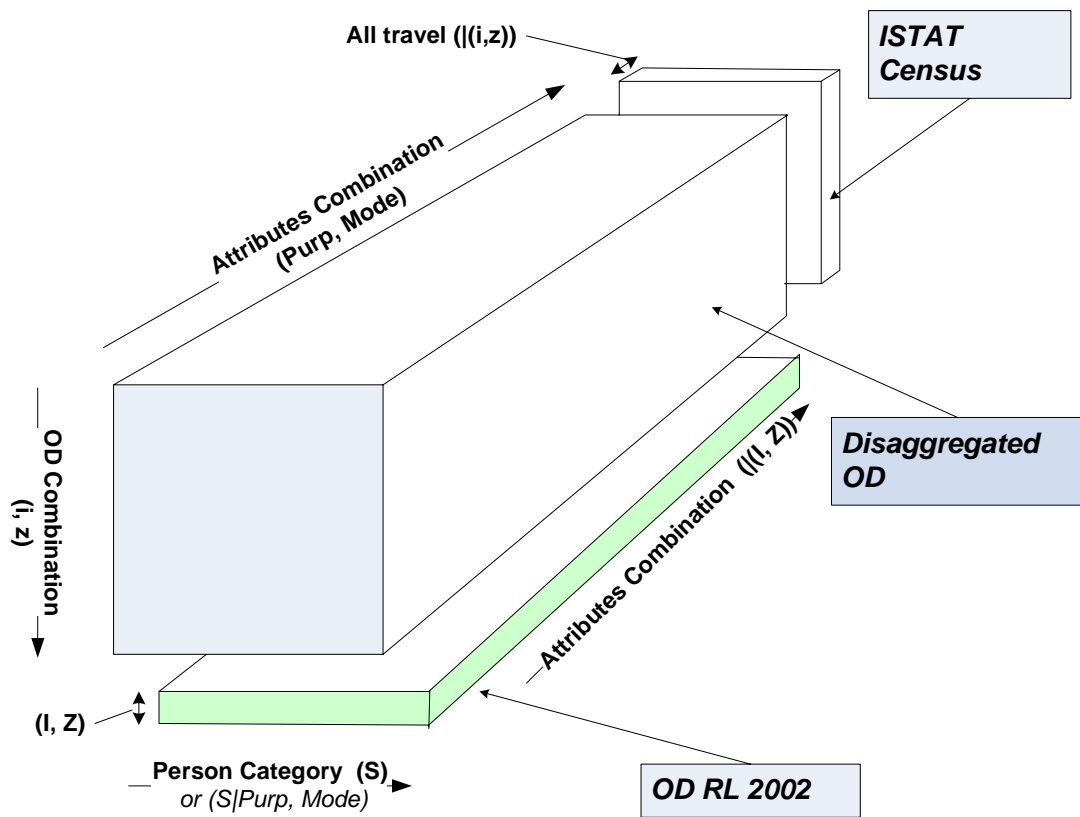


Figure 2.7 Multi-dimensional Data Cube

The dimensions of the various axes are marked. Thus, ‘Attributes Combination’ would, for example, be composed of a combined list of the form:

mode 1, purp 1

mode 1, purp 2

...

mode 2, purp 1

and so on.

The (horizontal) dimension of the Person Category dimension (S), which links between datasets, is based on either a simple list of Professional Category or, additionally, includes Age and Gender dimensions. The desirable alternative, derived from modelling, of having a Person Type dimension that contains the number of people by type given the travel purpose and mode (S|Purp, Mode) is also indicated.

2.7.2 Application of Dominici Method

For the purposes of this Lombardia feasibility study, the focus is on examining the extent to which the disaggregation problem can be cast in a form that has been shown to be operational by other parts of the OPUS project.

It is therefore necessary to determine the extent to which the specification in Section 2.7.1 meets the requirements of the Dominici method.

In this respect, the specification that has been considered so far might can be considered to allow the application of the Dominici method, but there is only limited information to with which to synthesise the disaggregate OD cube (as represented in Figure 2.7 above), so the uncertainty associated with the results will still be large.

It is therefore advisable to offer an improve basis for applying Dominici through providing an initial estimate of the disaggregate OD cube. This may be readily done through using a standard iterative proportional fitting (IPF) method having seeded the cube with non-zero values.

The application of the Dominici method then allows the distributional aspects of the observed data to provide a disaggregate OD cube that is better related to the statistical properties of the observations that would be available through IPF methods alone.

2.7.3 Scalability

Although the specification of the multi-dimensional data cube given in Section 2.7.1 is for a single ODRL zone combination, the dimensionality of the problem can still become large. As an example, the Person Type category might be defined to be composed of 5 Professional Categories, 3 Age Groups, and 2 Genders, giving a combined dimension of 30. The Travel Attributes category might have 6 modes and 5 travel purposes, giving another 30 dimensions. The OD combinations (that relate to Census zones, z , in ODRL zones, Z) depend on whether the focus is separately on trip generations and attractions, or on movements between zones. Assuming, say, 7 Census zones per ODRL zone on average, then this dimension could be 49, giving a combined dimension of $30 * 30 * 49$, which equals 44,100.

In practice, many of these dimension combinations are irrelevant, for example, rail trips for travel within the area shown in Figure 2.1 are not reasonable. Thus using reasoning of this type, it is possible to reduce the dimensionality considerably. The approach of using combinations dimensions, as discussed above, makes it easy to eliminate irrelevant combinations without detriment to the structure of the problem.

The OPUS project has demonstrated the application of the Dominici method using both MCMC Bayesian sampler and Maximum Likelihood methods [D7.3], with the latter offering faster computation times.

The issue of scalability is therefore a matter to be concerned about, but there are several avenues that are available for tackling this matter in application.

2.8 Traffic Flow Data

Part of the data available to the Lombardia region concerns traffic flows, which are available for links of the network. This is shown in the BBN diagram of Figure 2.6 as the final box for V_{ams} .

The use of such count data, together with path routing information from a transport model, has previously been reported [D10.2] and from which Figure 2.7 is reproduced.

The use of count data using the Tebaldi method [D4.2 Supplement] has been applied by OPUS for studies of Innsbruck (as a smaller test case) and for Zurich [D9.3], while the London study [D8.2] has demonstrated the transfer of TRIPS format network information (as applying for Lombardia) to the Visum network modelling system that has been used so far by OPUS. The alternative of a direct connection to TRIPS network modelling system is allowed by the system design though this would require some additional programming work.

The feasibility of using count data to further enhance the disaggregated OD cube following from the derivation of the disaggregated OD information is therefore confirmed.



Figure 2.8 Network Model and Link Flows for Milan (and Lombardia)

It would require that the detail of the network was enhanced in the area of interest and that the disaggregated zones (with centroid connectors) were introduced. This are procedural matters and do not pose any conceptual problems.

3. FEASIBILITY STUDY FOR BELGIUM

3.1 Problem Statement

For the Belgian case, as stated in D10.2, the feasibility study aims at investigating how the OPUS methodology could be used for a broader area, i.e. at a national level. The goal of the study is the determination of mobility profiles for the whole Belgian country. The first aim in this case is building a synthetic population for Belgium, at municipalities level (it means 589 entities) and then associating mobility profiles with it. The purpose of this step is twofold: first, estimating mobility indicators at rather disaggregated level and second (not treated here), building O/D matrices with the same granularity (589 X 589 cells). Moreover, a further step means coupling these estimations with demography models, allowing so forecasting mobility indicators for Belgium at a 30 years horizon. All these issues have already been partly addressed within national Belgian research programmes (SAMBA and MOBIDIC projects funded by the Belgian Science Policy within the framework of the “Scientific Support Plan for a Sustainable Development Policy II” programme). Nevertheless, the approaches followed during these researches were deterministic (e.g. least squares methods) and suffered from some drawbacks. We will remind the core of this deterministic approach in next subsection. Therefore, our plans within this feasibility study are to examine how the OPUS stochastic methodology could be more fruitful and more suitable for our Belgian case. Especially the OPUS methodology, as described in the OPUS documents D4.2 and D4.2 Supplement, seems adequate for taking into account jointly different data sets with their own uncertainties and to impute data for disaggregated entities relatively to known margin sums.

3.2 Previously used deterministic approach

Before going into details on how the OPUS methodology could be fruitfully used in our feasibility study, it seems worthwhile to sketch how we undertook our problem using deterministic techniques.

In order to understand our approach, let us recall the required goal: generating a population which satisfies some constraints (given by the margins known for the "true population") while remaining close to the observed population (through a sample). This problem can be formulated as a least-squares problem. This problem formulation has two advantages from our point of view: it can handle moderate inconsistencies in data sources and any under determinacy can be accommodated by classical regularization techniques.

Unfortunately, we had to give up building the whole Belgian population in one go because simultaneously creating synthetic populations for all districts at once leads to a least-squares problem too big to be practically manageable. We therefore build the synthetic population at the municipality level, but district by district, relying only on marginal sums derived from census or administrative datasets.

We now formulate this least-squares problem. To simplify the exposition, let us assume that the population is built by only considering three socio-demographic variables: age, gender and professional status (the method can be generalised without difficulty to more dimensions). Here is the available data and their notations:

- at the district level, we have the following cross-tables: (age x gender) and (gender x professional status);
- at the municipality level, we have margins for the various categories of the three variables but no cross-table.

Let us define the two following matrices (for each district)

- a matrix specific to a district a (age \times gender \times professional status) whose elements are noted α_{ijk} , with i for the age index, j for the gender index and k , for the professional status index;
- a matrix (age \times gender \times professional status \times commune) whose elements are noted c_{ijkc} , $c = 1, \dots, \#c$, c being the index of municipality and $\#c$, the number of municipalities in the district, the other indexes being defined as above.

We assume that each cell contains a number of individuals (belonging to a class of individual of age i , of gender j and of professional status k - in the commune c when it is necessary to specify it).

The objective is to estimate the four-dimensional matrix (age \times gender \times professional status \times commune), e_{ijkc} , given the following totals:

- known margins at the level of district a :

$$\alpha_{i.j.} = \sum_k \alpha_{ijk}$$

$$\alpha_{.jk.} = \sum_i \alpha_{ijk}$$

- known margins at the level of municipality c level:

$$T_{i..c} = \sum_{j,k} c_{ijkc}$$

$$T_{.j.c} = \sum_{i,k} c_{ijkc}$$

$$T_{..kc} = \sum_{i,j} c_{ijkc}$$

the least-squares problem is expressed as:

$$\min \sum_{i,j,k} \left(\sum_c e_{ijkc} \right)^2$$

$$\text{s.t.} \quad \begin{cases} \sum_{j,k} e_{ijkc} = T_{i..c} & \forall c \in a \\ \sum_{i,k} e_{ijkc} = T_{.j.c} & \forall c \in a \\ \sum_{i,j} e_{ijkc} = T_{..kc} & \forall c \in a \end{cases}$$

Municipality margins constraints

$$\begin{cases} \sum_{c \in a} (e_{i..c} + e_{.j.c}) = \alpha_{ij.} \\ \sum_{c \in a} (e_{.j.c} + e_{..kc}) = \alpha_{.jk.} \end{cases}$$

District margins constraints

For cross-tables (e.g. age \times gender) that are known at the municipality level, we get the following margins:

$$\gamma_{ij.c} = \sum_k c_{ijkc}$$

The constraints block related to these additional margins will then contain the new constraint:

$$(e_{i..c} + e_{.j.c}) = \gamma_{ij.c}$$

This problem has been encoded in the SIF (Standard Input Format) language as an input to the LANCELOT optimization package (Conn, A., Gould, N. and Toint, Ph. L. (1991) *LANCELOT a Fortran Package for Large Scale Non Linear Optimisation (release A)*).

Springer series in computational mathematics 17 Springer-Verlag (Heidelberg)) for its solution. Forty-three problem files (one for each Belgian district) have been created. Most problems are large.

This first formulation was however problematic because of the presence of linear dependencies in the constraints. These dependencies arise from the fact that some constraints are linear combinations of other constraints. Moreover, the removal of the redundant constraints revealed the fact that the problem was formally inadmissible because of the inconsistencies of our constraints. Consequently the least-squares problem has been modified, on one hand, for preventing the appearance of linear dependencies, and on the other hand, to allow moderate constraint inconsistencies.

We have thus considered an unconstrained weighted least-squares formulation instead of a constrained one. For each district a including c municipalities, we solve the following problem:

$$\begin{aligned} \min \quad & \sum_{i,j,k,c} (w_v e_{ijk})^2 + \sum_{c \in a} \left(w_{c_1} \left(\sum_{j,k} e_{ijk} - T_{i..c} \right) \right)^2 + \sum_{c \in a} \left(w_{c_2} \left(\sum_{i,k} e_{ijk} - T_{.j.c} \right) \right)^2 + \sum_{c \in a} \left(w_{c_3} \left(\sum_{i,j} e_{ijk} - T_{..kc} \right) \right)^2 \\ & + \left(w_{a_1} \left(\sum_c (e_{i..c} + e_{.j.c}) - \alpha_{ij.a} \right) \right)^2 + \left(w_{a_2} \left(\sum_c (e_{.j.c} + e_{..kc}) - \alpha_{.jka} \right) \right)^2 \\ \text{s.t.} \quad & e_{ijk} \geq 0 \end{aligned}$$

Weights related to each kind of constraints are respectively written $w_v, w_{c_1}, w_{c_2}, w_{c_3}, w_{a_1}$ and w_{a_2} .

The 43 weighted least-squares problem could be solved by LANCELOT.

However, predicting mobility indicators for the considered horizon requests further data analysis, linking the spatial and socio-demographic population description with these indicators. This link can be obtained from travel behaviour survey like MOBEL (Hubert, J.-P. and Toint, Ph. L. (2002) *La Mobilité quotidienne des Belges*. Presses Universitaires de Namur (Namur)), at least for the year 2000 and after further data processing. If we define a population group to be the set of all Belgian residents sharing the same age, gender, household type, education level, activity status and licence ownership (that is all the P-class characteristics except the localization), we are able to derive from MOBEL a set of mobility indicators (mobility rates, modal shares, travel purposes, ...) for each observed such group. Unfortunately, the survey sample does not contain a significant representation for each population group, and the indicators of interest are only available for a (sparse) subset of these groups, which may itself depend on the indicator considered. We therefore designed a completion process for the population groups, whose objective is to derive suitable indicators for all groups. This process is inspired by automatic imputation techniques such as the "hot deck" (see Lothaire, O. and Toint, Ph. L. (2003) *A Toolbox Approach to Data Correction and Imputation in Capturing Long-Distance Travel*, Axhausen, K.W., Madre, J.-L., Polak, J.W. and Toint, Ph. L. editors, Research Sciences Press (Baldock), pp.244-255) for a review of automatic imputation methods). Broadly speaking, it consists of attributing to individuals of a population group not represented in MOBEL a mobility behaviour determined by observed neighbouring groups, where neighbouring relations are measured in the 6-dimensional space of age, gender, household type, education level, activity status and licence ownership.

Once the desired mobility indicators are derived for all populations groups, they can be quantified for the reference years, in terms of localization (districts) and group size by considering the P-classes for these years. More specifically, each P-class for a given year then specifies, for that year, the district and size of each population group in the district, to which specific mobility indicators can then be attached. Aggregating the district population finally yields an estimate of the mobility indicators for each district and for each reference year.

Then the forecasts based on the population's evolutions (D-classes) provided by demography models can be computed the same way for the future mobility indicators.

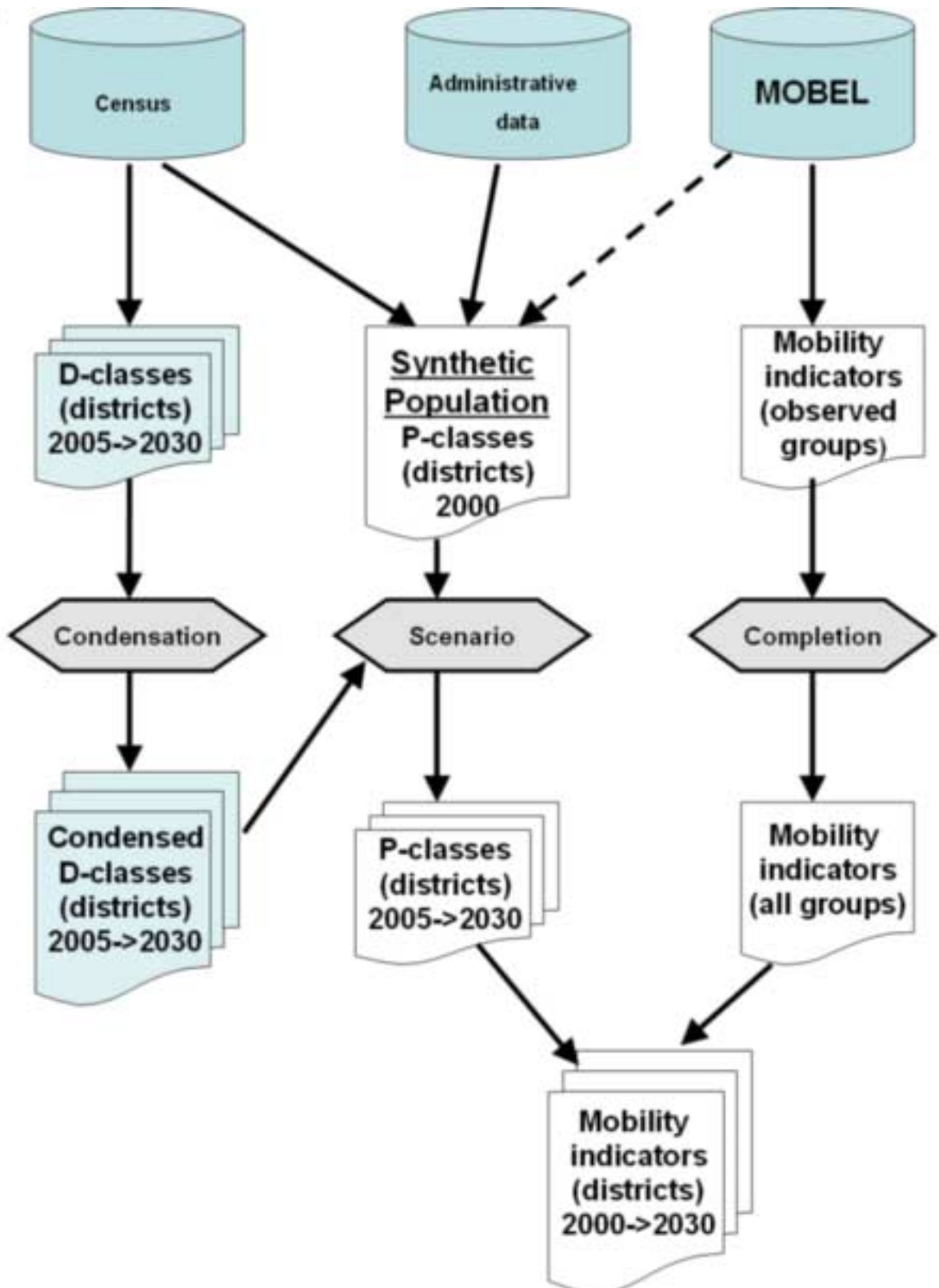


Figure 3.1. Deterministic processes for associating mobility indicators to synthetic population and generating forecasts (Belgian feasibility study)

3.3 Applying the OPUS Methodology

The core of this deliverable is now to describe how the OPUS methodology could be used instead of our deterministic least squares method. This application means following several steps. The first one starts from the problem specification and from the available data and converts them into a set of conceptual abstractions that generate a holistic view of the system being considered. A GAPM (Generalised A Priori Model) is the general tool used for this transformation. The next step goes from these abstractions to a more formally formulated and more precise description. The result provided by this phase is a Bayesian Belief Network (BBN) formalizing our Belgian feasibility study. Finally we will also deal with contingency tables, tools allowing us to describe how we have to estimate disaggregated cells from margin totals (in this Belgian feasibility study, the disaggregated cells represent the number of people in each detailed (by gender, age class, diploma class, etc) category of the population in a municipality and the margin totals are the number of people in less detailed categories (e.g. all the women together or all the people between 18 and 30) or at a larger spatial level (i.e. the district level in our Belgian case)).

3.4 The GAPM

The OPUS methodology starts from the derivation of a GAPM (Generalised A Priori Model) which graphically shows the main components for the studied case, here the Belgian mobility profiles for detailed population groups at municipalities level, and how they interact. In this formalism, the data and models sources are mainly the core of the represented components.

Regarding the Belgian feasibility study, the GAPM, at a macro level, is quite simple since few main components are taken into account.

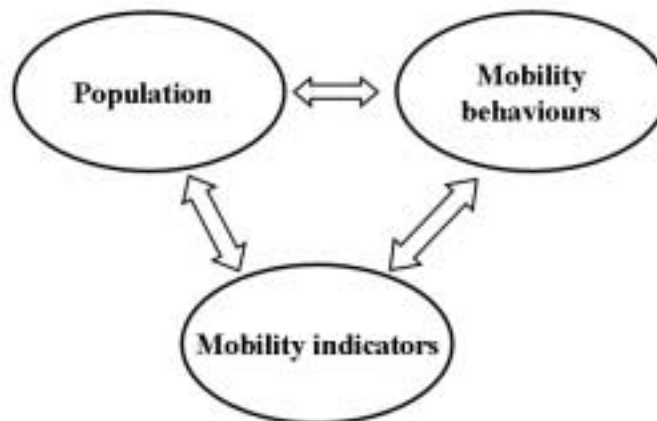


Figure 3.2 Main components for Belgian feasibility study GAPM

The GAPM from Figure 3.2 shows that we plan to mix information about the population with data about their mobility behaviours in order to obtain mobility indicators. The main issue is that these results have to be provided at a quite disaggregate spatial level (municipalities) taken also into account a detailed categorization of the population where classes are defined by gender, age ranges, diploma, type of household, professional status and driving license ownership.

Thus, we can describe these three main components more precisely.

The **population** box corresponds to a fine description of the Belgian population characteristics. As well demography factors as gender, age and type of households as socio-economic parameters like diploma, professional status (active, inactive or student) and driving

license ownership define the categorization of the population. Moreover a spatial disaggregation is also taken into account since these categories are localized at municipality level. Some of the elements playing a role in this characterization of the population are drawn from the National Register (for more details see the OPUS deliverable D10.2) [gender, age, household type], other from censuses (2001 census if available, otherwise 1991 census) [professional status, diploma] and last one from other administrative data sets (driving licenses database from the Ministry of Transports and Mobility). For privacy respect, all these data are provided after some aggregation process and therefore only margin sums (at different levels and for different “dimensions”) are available.

The **mobility behaviour** box corresponds to the information collected during the MOBEL survey. Linked with the demographic-socio-economic characteristics of the respondents, this data set contains the description of their displacements during a given day. All this information is fully disaggregated.

Finally, the **mobility indicators** box will provide the results about some indicators like the modal share, the travelled distance of the departure time (peak or off peak) distribution associated with the population of each municipality. These results are computed by associating the information contained in the two previous components.

Now, all these details can be incorporated in a more elaborate GAPM which is given in Figure 3.3.

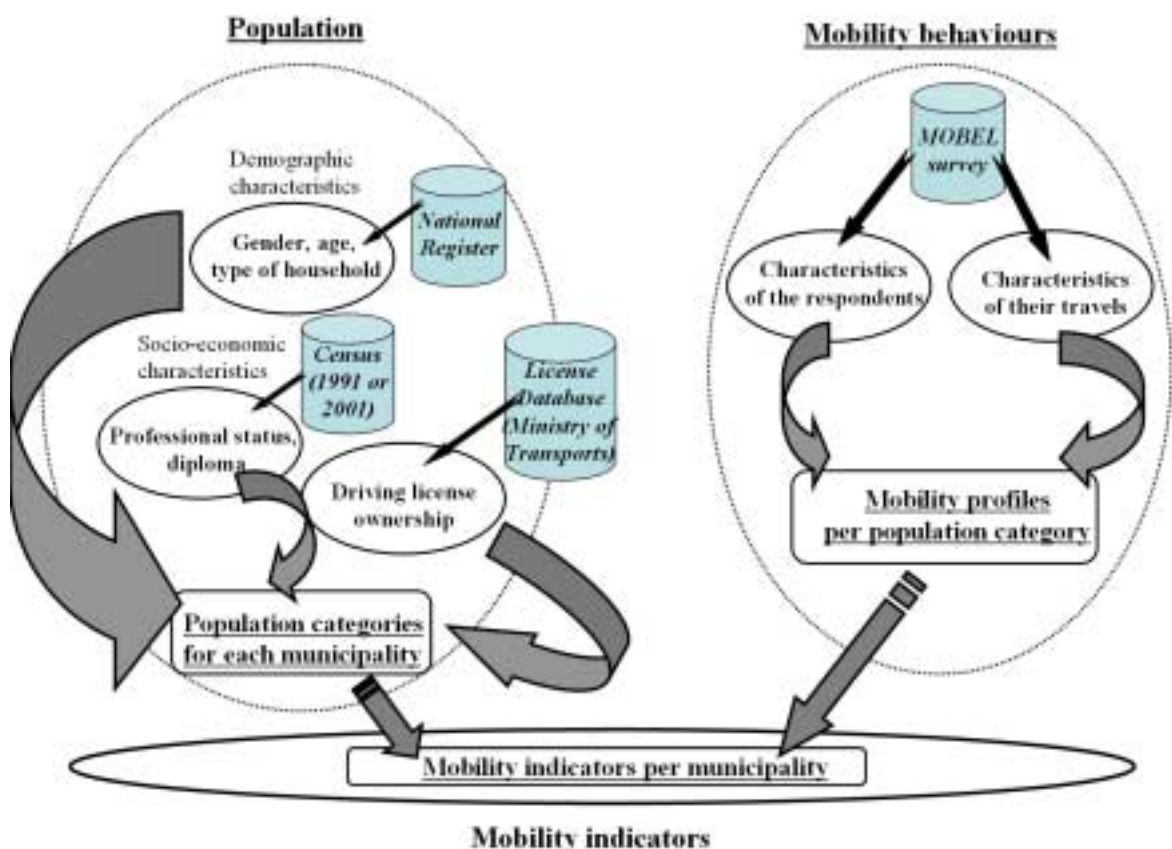


Figure 3.3. : Detailed GAPM for Belgian feasibility study

3.5 Sources of Data

From the different data sources described in the OPUS deliverable D10.2 and represented in the GAPM of Figure 3.3, the following information was extracted for generating the

conditional probability distributions to be provided for the Bayesian Belief Network description.

3.5.1 For the population:

We draw different margin sums representing the number of people per

- Municipality \times age \times gender (from National Register)
- Municipality \times age \times gender \times driving licence (from driving licenses database of the Ministry of Transports and Mobility)
- Municipality \times type of household (from 2001 census data)
- Municipality \times professional status (from 1991 census data)
- Municipality \times diploma (from 1991 census data)
- District \times type of household \times age (from 2001 census data)
- District \times gender \times professional status (from 1991 census data)
- District \times age \times professional status (from 1991 census data)
- District \times gender \times diploma (from 1991 census data)
- District \times professional status \times diploma (from 1991 census data)
- District \times age \times diploma (from 1991 census data)

3.5.2 For the mobility behaviours:

From the MOBEL database, we draw

- The same demographico-socio-economics characteristics as the ones defining the categories in the population : age, gender, type of household, professional status, diploma and driving license for each respondent
- Some indicators describing the trips of the “mobile” (i.e. the ones having at least achieved one trip during the given reference day) respondents : mode, departure time, distance, ...

3.6 The BBN

Next step, according to the OPUS methodology, consists in transforming the GAPM (from Figure 3.3) into a BBN where the mathematical relationships between the elements have to be clearly defined.

The BBN is formed by following the flow of interactions expressed in the GAPM (Figure 3.3) and forming it into a directed graph.

For the population part, we start with N the number of people in the Belgian population and its share in municipalities population (N_c), our goal is to compute $P(s, a, h, p, d, l | c)$ meaning the probability to be of gender s , in age class a , in a household of type h , with a professional status p , a diploma d and a driving license ownership status l (i.e. having or not a driving license) when living in a municipality c .

We dispose of a series of information for calculating these probabilities:

- $P(s, a | c)$ i.e. the probability of being of gender s and age class a when living in municipality c
- $P(s, a, l_1 | c)$ i.e. the probability of being of gender s and age class a and having a driving license (l_1) when living in municipality c

- $P(h|c)$ i.e. the probability of belonging to a household of type h when living in municipality c
- $P(p|c)$ i.e. the probability of having a professional status p when living in municipality c
- $P(d|c)$ i.e. the probability of having a diploma of type d when living in municipality c
- $P(a, h|D)$ i.e. the probability of being in age class a and belonging to a household of type h when living in district D
- $P(s, p|D)$ i.e. the probability of being of gender s and having a professional status p when living in district D
- $P(a, p|D)$ i.e. the probability of being in age class a and having a professional status p when living in district D
- $P(s, d|D)$ i.e. the probability of being of gender s and having a diploma of type d when living in district D
- $P(p, d|D)$ i.e. the probability of having a professional status p and a diploma of type d when living in district D
- $P(a, d|D)$ i.e. the probability of being in age class a and having a diploma of type d when living in district D

Stricto sensu, all these probabilities are given by the corresponding ratios between the margin sums and the total populations (e.g. $P(s, a|c) = \frac{n_{sac}}{N_c}$)

The result we have to computed ($P(s, a, h, p, d, l|c)$) could be related to all these other probabilities by equations like :

$$\sum_h \sum_p \sum_d \sum_l P(s, a, h, p, d, l|c) = P(s, a|c)$$

or

$$\sum_{c \in D} \sum_s \sum_h \sum_p \sum_l P(s, a, h, p, d, l|c) = P(a, d|D)$$

The final results will be the $n_{sahpdlc}$ being calculated as

$$n_{sahpdlc} = N_c \cdot P(s, a, h, p, d, l|c) \quad \forall s, a, h, p, f, l, c$$

For the mobility behaviours part, we will focus here on the modal split. The same process could be followed and applied for other indicators like the travelled distance or the departure hour.

Thus, the goal is here to provide $P(m|s, a, h, p, d, l)$ i.e. the probability of using mode m when being in the category of the population characterized by gender s , age class a , household type h , professional status p , diploma d and driving license ownership status l .

These probabilities have to be derived from collected data (from the MOBEL survey). Thus they are straightforwardly computed as

$$P(m|s, a, h, p, d, l) = \frac{t_{sahpdl}^m}{t_{sahpdl}}$$

where t_{sahpdl}^m is the observed number of trips travelled with mode m by people from population category “sahpdl” and t_{sahpdl} is the total observed trips number for the same population category.

Evidently, this equation is only valid if t_{sahpdl} is different from 0 (or more pragmatically over a given threshold of t observations (e.g. $t = 10$). Otherwise, specific imputations techniques must be undertaken.

Finally, the mobility indicators per municipality are computed as (here once more given for the case of the modal share)

$$P(m|c) = \sum_s \sum_a \sum_h \sum_p \sum_d \sum_l n_{sahpdlc} P(m|s, a, h, p, d, l)$$

Therefore the corresponding BBN can be drawn as shown in Figure 3.4.

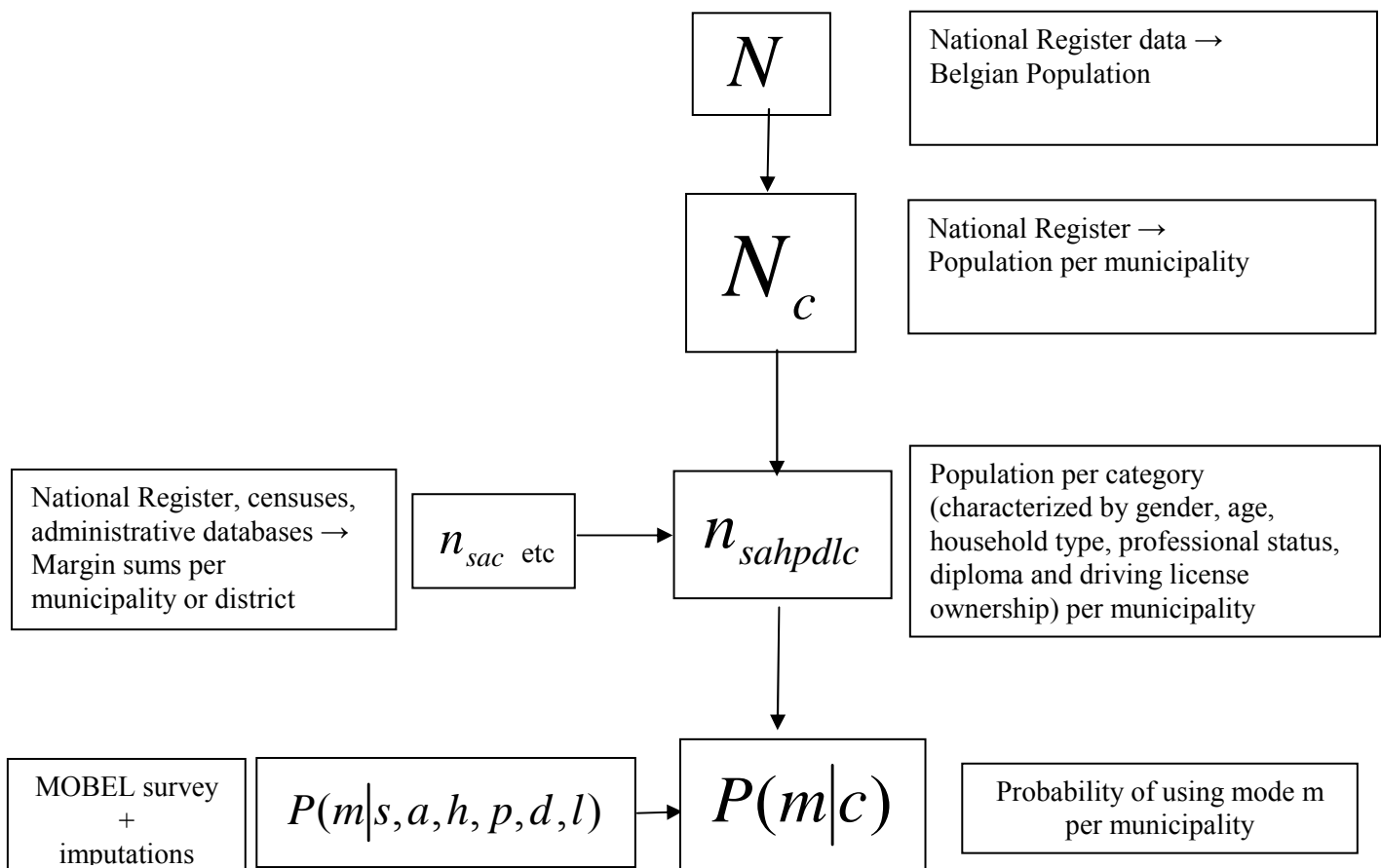


Figure 3.4 Bayesian Belief Network (BBN) for Belgian feasibility study

3.7 Application of Dominici Method

The main computation processes to be achieved within the Belgian feasibility study could be both considered as estimating cells values within a multidimensional matrix for which margin sums (in some dimensions) are known.

These multidimensional matrices could also be called contingency tables.

Such situations are encountered when

- We compute $P(s, a, h, p, d, l|c)$ knowing margin sums $P(s, a|c)$, $P(s, a, l_1|c)$, $P(h|c)$, $P(p|c)$, $P(d|c)$, $P(a, h|D)$, $P(s, p|D)$, $P(a, p|D)$, $P(s, d|D)$, $P(p, d|D)$ and $P(a, d|D)$.
- We have to impute $P(m|s, a, h, p, d, l)$ because t_{sahpdl} is under a given threshold. In such a case, we can suppose we know some margin sums like $P(m|s, a, h, p, l)$ or $P(m|s, a, d, l)$ or ...

These two cases are slightly different: in the first one, all the cells are to be estimated knowing the margin sums; in the second, some cell values are already known from the information collected during the MOBEL survey whilst other ones are to be estimated from the margin sums also deduced from the survey.

The Dominici method described in the OPUS document D4.2 Supplement could be fruitfully used for such computations.

It must be mentioned that in the case of the computations of the $P(s, a, h, p, d, l|c)$ cells, the used margin sums are drawn from different data sets (National Register, driving licenses database, census 1991 or census 2001) and therefore present different uncertainties. This is also a reason why applying the OPUS methodology and the Dominici method could be a good opportunity to exploit their capabilities for dealing with such different data sets.

Nevertheless, the scalability issue, as already discussed for the Lombardia Regione, also plays an important case in this Belgian feasibility study. As some margin sums are only known at district level, we must treat together all the municipalities in one district, this mean an average of ± 15 municipalities. Moreover, in each municipality, the population categories are characterized by

- 2 genders
- 4 age classes
- 4 household types
- 3 professional statuses
- 4 diploma types
- 2 driving license ownership statuses

This means $2 \times 4 \times 4 \times 3 \times 4 \times 2 = 768$ categories per municipality. Even if this amount could be reduced to 536 categories, simply making following realistic assumptions:

- Individuals of less than 18 years do not own a driving licence.
- The age group of the 6-17 years only contains students.
- There are no student aged of 60 years or more.
- Individuals of less than 18 years do not have superior education diplomas,

the cells to be computed in one step are about 8000 ($536 \times \pm 15$).

This proves that the concern in the implementation of Dominici method taken into account during the OPUS project (see the OPUS deliverable D7.3) with particular focus on scalability issues addresses a really actual problem encountered in many “real life” transport situations.

4. SUMMARY AND CONCLUSIONS

4.1 Regione Lombardia Feasibility Study

This study has examined how the OPUS methodology might be used to take advantage of the regional travel survey (ODRL) for application at a smaller spatial scale, for which a small part of Milan has been considered for the purposes of investigation.

The Feasibility study has described the manner in which the ODRL and ISTAT Census data could be used to provide the basis of a disaggregation procedure that exploited the statistical details of these two large datasets. The results are further enhanced by the use of traffic count data.

The system that is described might be regarded as being rather minimal in nature but, importantly, provides a framework in which data from local surveys might be incorporated.

It was determined [D10.2] that local surveys (for example, trip destination surveys) were not available as might be desired, but the methodology enhances the value of undertaking such surveys and so, in principle, makes it more likely that they would be undertaken.

The Feasibility study has described a method of ‘dimension combinations’ that is sufficiently general that further attributes (from additional) surveys may readily be admitted into the procedure. However, the study has provided some warnings about the scalability of the method when the number of the dimensions is extended and this issue must be monitored carefully in any applications that might be considered for Regione Lombardia.

4.2 Belgian Feasibility Study

The applicability of the OPUS methodology on a problem covering whole a country has been evaluated on the estimation of disaggregated, both spatially and on other demographico-socio-economic parameters, mobility indicators for Belgium.

Moreover, this Belgian feasibility study has described how the computation process has to deal with data drawn from different data sources, each with its own uncertainties.

The Belgian case has been formulated through the different steps (GAPM, BBN) followed by the OPUS methodology and it has been demonstrated that this problem could be suited to the OPUS methodology.

Finally, it was described how the Dominici method, implemented within the OPUS project, could be a useful tool for the disaggregated estimations to be undertaken in the Belgian feasibility study.