OPUS

Optimising the use of Partial information in Urban and regional Systems

Project IST-2001-32471

WP11: Feasibility Study - HEALTH

Title :	**Feasibility Study Design**

Author(s) :	Marco Martuzzi	World Health Organization
	Francesco Mitis	World Health Organization
	James Bennett	DEPH, Imperial College
	Maria De Iorio	DEPH, Imperial College
	Nicky Best	DEPH, Imperial College

Deliverable No. : D11.2
Version : 1.6

Contract Date : November 2005
Submission Date : December 2005

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 : 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 D11.2 is the second part of Work Package WP11 of the OPUS project. Work Package WP11 is titled: “Feasibility Study - Health”. It explores the applicability of the methods developed by the OPUS Consortium to a second domain, specifically that of health.

The focus of the work is on identifying issues which arise when applying the statistical methods or database technologies to health sector problems.

Deliverables of WP11 are divided in two sections. In a first part, developed by the World Health Organization, tools for a practical application of health impact assessment methods, built to estimate burden from environmental hazards, will be provided and potentially included in the OPUS methodology; in the second section, developed by the Department of Epidemiology and Public Health (DEPH), Imperial College, the integration of different sources of information, including daily activity patterns in the area of exposure modelling, will constitute an example of the use of OPUS methodology in epidemiology.

1. INTRODUCTION AND OBJECTIVES

1.1 OPUS Project Work Package 11

1.1.1 Objectives

The objectives of Work Package 11 (WP11) are:

- Explore the applicability of the model developed to a second domain, specifically that of health.
- Identify any new issues either for the statistical methods or for the database technologies that are raised by the problems from the health sector.
- Develop a plan for the implementation of the methods in the health sector.

1.1.2 Description of WP11 Work

World Health Organization, European Centre for Health and Environment (WHO) and Imperial College, Department of Epidemiology and Public Health (DEPH) will develop a feasibility study focussed on the health impact of socio-economic and environmental hazards. The work will be divided in two sections.

WHO will introduce deterministic health impact assessment (HIA) methods. As an example of application of these methods, particulate matter concentration in urban areas will be considered as the main environmental hazard and HIA methods to estimate its health effects will be developed. Problems emerging from different socio-economic gradient of exposed population will be included in the analyses, too.

The three deliverables will be structured in the following way: in the first D11.1 a review of the health impact assessment methods, based on deterministic relations, will be described; in the deliverables D11.2 a Bayesian model will be implemented and all the components regarding every source of uncertainty in this kind of studies will be stressed: this model will be incorporated in OPUS methodology; the third deliverable D11.3 will be formed by a review of the wide range of health effects that can be potentially estimated.

DEPH will review small area health study methods and aim to gain a better representation of area level exposure than is routinely used. The feasibility of these methods will be explored in an example relating hospital admissions and deaths for cardio-respiratory disease to NO_2 exposure. Issues arising from the combination of two independent surveys will be explored. A further potentially interesting topic is the joint modelling of pollution surfaces collected at different levels of spatial resolution.

The deliverables will be presented in the following way: in D11.1 typical small area health studies will be reviewed, considering those areas in which improvement might be made; in D11.2 details of the available data sets will be given, together with a Generalised A Priori Model (GAPM) and Bayesian Belief Network (BBN); D11.3 will include a review of the contributions made by the OPUS methodology together with the code required for implementation of such a model.

1.2 Structure of the Deliverable D11.2

In the second section, the World Health Organization has introduced a non-deterministic model for the calculation of health effects deriving from air pollution exposure. General issues about Health Impact Assessment and air pollution have been described in § 2.1; some differences between a deterministic and a stochastic approach have been described in § 2.2; the Bayesian model, with graphical structure and model code, has been introduced in § 2.3 and § 2.4; some selected results regarding an example of health impact assessment of PM_{10} in the eight major Italian cities have been reported in § 2.5 to illustrate the path that has been followed to switch from a deterministic to a probabilistic approach; the characteristics and role of uncertainty in health impact assessment models used to estimate the attributable burden of disease due to PM_{10} exposure has been described in § 2.6. With few modifications, substantially regarding the age-groups to be considered, the Bayesian model that has been presented can be used for a wide set of adverse health outcomes that will be analyzed and described in deliverable D11.3.

In the third section DEPH introduce an exposure model based on the integration of data from two surveys. In § 3.1 the motivation for this exposure model is given, together with a graphical representation of the overall disease model; § 3.2 describes the data sets which we make use of in the modelling; in § 3.3 the way in which the elements of the exposure model will combine and the statistical model for transport activity are described. A graphical model of this model is given in § 3.3.1. In § 3.4 the WinBugs code used for implementing the model is given. Finally in § 3.5 there is an outline of the ongoing work on the model.

2. A BAYESIAN HEALTH IMPACT ASSESSMENT MODEL

2.1 Health Impact Assessment

Health Impact assessment (HIA) is of growing interest in the field of public health in Europe. Scientific knowledge on the adverse effects of several environmental factors on the population is in some cases substantial, but often-regulatory policies fail to reflect such knowledge adequately. In addition the Amsterdam Treaty of the European Union (Article 152) states that definition and implementation of all new Community policies should ensure a high level of human health protection, which can be achieved through HIA.

A key element to support the formulation of public health policies based on the available evidence is to develop rigorous methods to translate experimental, toxicological and epidemiological information into accurate estimation of the overall impact on the health of the population. So far, most of these exercises have been based on relatively simplistic methods (Kunzli et al., 1999, Martuzzi et al., 2002, Martuzzi, Krzyzanowski & Bertollini, 2003), as showed in D11.1, where the health impact is measured by direct derivation of attributable risks, based on available concentration-response estimates and exposure profile of the population (often an average measure), and not using statistical models.

Uncertainties in the estimates of the impact are normally evaluated on the basis of the confidence intervals attached to the dose-response function, and other sources of uncertainty that can be of great importance are ignored, or acknowledged only qualitatively. In addition, concerns of “double counting” of events in multiple health endpoints assessments, often suggest limiting studies to a partial set of adverse health consequences. As a result, impact assessments are often said to be “conservative”, and little or no attempt is made to verify that the necessary assumptions are met and/or to carry out quantitative evaluation of the true systematic and random error involved.

In addition to that, the attention towards estimates dealing with the same problems in terms of potential years of life lost has been growing (Hurley, J. F. et al., 2000, Murray & Lopez, 1996, Rabl, 2003, Roosli et al., 2005), in order to identify more specific and sensible population groups and develop a more structured approach in public health policies so that targets can be established and goals achieved.

Using data on urban air pollution and health, estimates of the effects of shifting risk factor distributions towards a counterfactual (or referent) factor, rather than the difference between “exposed” and “unexposed”, have been calculated (Ostro, 2004). “Conventional” statistical methods, both for a short-term and for a long-term approach, have been explored in D11.1. A Bayesian methodology (for a short-term approach) including the calculation of years of life lost, has been illustrated in this deliverable D11.2. This approach leads to the calculation of the attributable burden in a given population due to a specific exposure (PM_{10} ambient concentration in our case), and, consequently, to the portion of disease burden (the avoidable burden) that could be reduced in the exposed population if causative exposure were eliminated, taking into account more sources of uncertainties. The most important of them is the variability of the city-specific distribution of PM_{10} concentrations: not only yearly average values have been taken into account for the estimation of health effects. This heavily conditions the precision of the final estimates.

2.2 A Bayesian Model: introduction

Normally, the use of conventional methods allows including in the final results only the uncertainty attached to the epidemiological concentration-response estimates (RRs), most of times based on a simple linear relation. To consider more sources of uncertainty, such as the variability attached to the distribution of PM_{10} concentrations, variability in the exposure and mortality, uncertainties in the use of not equivalent metrics among different fixed-site monitoring stations it is necessary to switch from a deterministic approach to a probabilistic and hierarchical model.

To improve the reliability of the uncertainty estimates, to take into account missing data, to solve multivariate statistical problems for which it can be difficult to find an exact analytical solution without the use of iterative techniques, to create latent (not observable) variables, alternative statistical methods based on a Bayesian approach have been explored.

A Bayesian probability model, in which most of the quantities are stochastic variables, has been defined and applied to the data to explore the relationships between the distribution of air pollutants (PM_{10}) and health effects.

The model consists of a full joint probability distribution of all the unobserved (i.e. parameters and missing values) and observed quantities (demographic, health and environmental data). This distribution is conditioned by data and posterior distributions for parameters and unobserved quantities are obtained. From each of these *h* conditional posterior distributions, alternatively for each unknown parameter and keeping fixed the remaining ones, one value is generated through a large number of simulations. It has been demonstrated that, under general conditions, every set of simulations (chain) can be considered as a sample extracted from the joint distribution of probability (Richardson & Gilks, 1993).

Several statistical methods to study the convergence of the simulations to the true value have been developed in the recent years. The Gelman-Rubin method (Gelman & Rubin, 1992), as modified by Brooks and Gelman (Brooks & Gelman, 1998), has been applied. It is a convergence test based on two or more parallel chains, each started from different initial values which are over-dispersed with respect to the true posterior distribution: it is based on a comparison of the within and between chain variances for each variable (essentially a classical analysis of variance). Convergence is diagnosed when the chains have “forgotten” their initial values, and the output from all chains is indistinguishable

After the convergence between two or more chains of simulated values has been checked and a further number of iterations has been made to decrease the overall variability, the estimates for the unknown parameters are derived from the joint posterior distribution through the use of some summary measures, descriptive statistics as means, medians and percentiles (Tierney, 1983). In addition to that, inference can be made on any function of the estimated parameters. To get the final estimates a sufficient number of simulations (in our case, 10 000 for each chain, without considering in the analysis the first 1 000, needed to reach the convergence and then discarded) has been done through the Monte Carlo method described above using the WinBUGS software (Spiegelhalter, Thomas & Best, 1999).

The algorithm used to estimate the unobserved quantities belongs to the family of Monte Carlo simulation iterative methods and it is a procedure of numerical integration known as Gibbs sampling (Gelfand, Hills & Racine-Poon, 1990).

Through this Monte Carlo simulation method it is possible to calculate attributable risks and number of cases with 95% credibility intervals. An example of health impact assessment of PM_{10} in the eight major Italian Cities have been reported to show the progresses obtained in the model, switching from a conventional to a probabilistic approach (§ 2.5). Effects have been calculated for every single city and for a latent variable, not observable, a joint variable “sum” of all the cities.

By disaggregating the outcome by sex and age-class attributable number of deaths for every city, every age-class (from 30 to >95 years) and sex have been obtained. Calculations have been made using different counterfactual, or referent, factors (20, 30 and 40 $\mu g/m^3$) for PM_{10} ambient concentrations. The algorithm used for the calculation of adverse health effects is the same presented in D11.1.

By using this Bayesian approach a more accurate assessment of the uncertainty in the data is possible: the width of the 95% credibility intervals also depends on the variability of the eight cities empirical distributions of pollutants and not only from the variability of the relative risk obtained from the meta-analysis. In addition to that, a joint variable, with its own variability, can be calculated for each sex.

To get the final values for every single city and for the total of the cities a sufficient number of simulations (10 000, after the convergence, for each chain) is done through the WinBugs software.

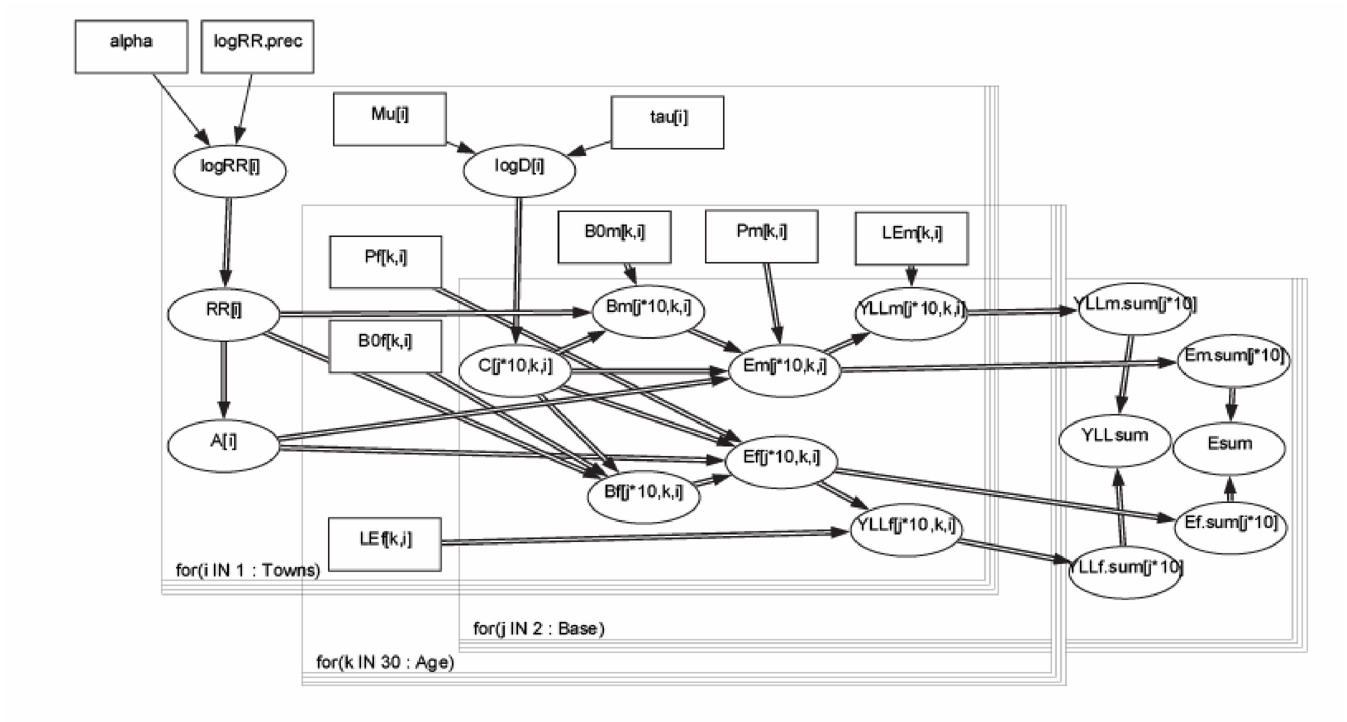
Details of the models, and procedures followed for model fitting are illustrated in § 2.3 - § 2.4.

2.3 A Bayesian Model: the graphical structure

The Doodle created through WinBugs 1.4, useful to better illustrate the logical structure of the model is attached below.

Ellipses represent variables that are given a distribution and in the hierarchical model can be “children”, if they depend on other variables, or “parents”, if other variables depend on them; they can be observed value (data) or unobserved ones (parameters). Rectangles denote constant values that cannot have parents, are fixed by the design of the study and must be specified in a data file.

The big rectangular boxes (called plates) represent cycle of operations to be executed for different set of variables (e.g. for different counterfactual exposure (j), towns (i); and age-classes (k). Links between variables can be of two types: a hollow arrow represents mathematical relations between two variables while a solid one indicates stochastic dependence.



2.4 A Bayesian Model: the model code

The model code, written in WinBugs language, with explanatory notes, is attached below.

model;

```
{
#To define the precision (the inverse of the variance) of the RR distribution, variability was taken from a WHO study (Martuzzi et al., 2002):
logRR.prec <- 1/(0.0084*0.0084)
```

```
#To introduce the Relative Risk variability and initialize the cycle for every single city "i" (in this case i from 1 to N, N=8 towns):
for( i in 1 : Towns ) {
A[i] <- (RR[i] - 1) / RR[i]
RR[i] <- exp(logRR[i])
logRR[i] ~ dnorm(0.0257,logRR.prec)
#where 0.0257 is equal to ln(RR=1.026), our RR

#To define the distribution shape of every single city:
logD[i] ~ dnorm(Mu[i],tau[i])
tau[i]<-1/var[i]}

#To initialize the cycle for every single city "i", for every age "k" from 30 to 95+, for every counterfactual factor "j"(from 2 to 4)*10 (20,30 and 40 μg/m^3):
for( k in 1 : Age ){
        for( j in 2 : Base ){
                for( i in 1 : Towns ){
YLLm[j * 10 ,k, i] <- (Em[j * 10 , k , i]) * LEm[k , i]
YLLf[j * 10 ,k, i] <- (Ef[j * 10 , k , i]) * LEf[k , i]
Em[j * 10 , k , i] <- (((A[i] * Bm[10 * j , k , i]) * C[10 * j , k , i]) / 10) * Pm[k , i]
Ef[j * 10 , k , i] <- (((A[i] * Bf[10 * j , k , i]) * C[10 * j , k , i]) / 10) * Pf[k , i]
Bm[j * 10 , k , i] <- B0m[k , i] / (1 + (RR[i] - 1) * (C[j * 10 , k , i] / 10))
Bf[j * 10 , k , i] <- B0f[k , i] / (1 + (RR[i] - 1) * (C[j * 10 , k , i] / 10))
C[j * 10 ,k, i] <- exp(logD[i])-10*j
}}}

#Where:
# Em and Ef are the number of attributable deaths, LEm and LEf are the residual life expectancies, YLLm and YLLf are the years of life lost: all these pairs of variables have been estimated separately for males and females, disaggregated by counterfactual factor, age and town;

#To define latent variables for different counterfactual factor, for the eight cities and for each sex and for all the variables together:
for( j in 2 : Base ) {
YLLf.sum[j*10]<-sum(YLLf[j*10, , ])
YLLm.sum[j*10]<-sum(YLLm[j*10, , ])
```

```
Ef.sum[j * 10] <- sum(Ef[j * 10 ,  , ])
Em.sum[j * 10] <- sum(Em[j * 10 ,  , ])
YLLsum[j]<- YLLm.sum[j * 10] + YLLf.sum[j * 10]
Esum[j] <- Em.sum[j * 10] + Ef.sum[j * 10]
}}
#Data section for 8 cities, 3 baselines, 66 age classes, with: 6 matrices of dimension 66 (ages) * 8 (cities) filled with population data, #mortality data and life expectancy data, 3 for each sex, 8 element vectors for average and standard deviations of the PM10 #distributions.
```

Data

```
List     (Towns=8,Base=4,Age=66,
Pm = structure(.Data =c(…).Dim = c(66,8)),
Pf = structure(.Data = c(…).Dim = c(66,8)),
B0m=structure(.Data=c(…).Dim = c(66,8)),
B0f=structure(.Data=c(…).Dim = c(66,8)),
LEm=structure(.Data=c(…).Dim = c(66,8)),
LEf=structure(.Data=c(…).Dim = c(66,8)),
Mu=c(3.870648262,3.781789932,3.712394,3.704094,3.682932,3.828847,3.825548,3.715791)
var=c(0.19239,0.013224,0.253578,0.400475,0.157141,0.234381,0.220987,0.125717))

#Two sets of widely different initial values for stochastic "nodes" connected to the two sources of variability: the relative risk RR and the pollutants concentration. They are used to initiate two chains of simulated values
```

Inits

```
list(logD = c(1,1,1,1,1,1,1,1),
logRR = c(0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001))
list(logD = c(2,2,2,2,2,2,2,2),
logRR = c(0.003,0.003,0.003,0.003,0.003,0.003,0.003,0.003))


#N represents the number of the cities, B0m and B0f the mortality rate values (66*8 values), Pm and Pf the vector of the exposed populations (66*8 values), Mu the vector of the 8 average values for each city, synthetic indicators of PM10 values, Tau the 8 precision vectors.
#Alpha and logrr.prec are constants defined outside the data section and they are equal for all the cities
```

2.5 The path from a deterministic to a probabilistic approach: selected results

Several models were used to explore some of the available options, as follows.

Model 1 (Conventional Model): Former WHO study results (Martuzzi et al., 2002) updated with population historical data - variability of Relative Risk; use of arithmetical mean;

Model 2 (Bayesian, First version): WinBugs, first version - same variability of Model 1 but Bayesian; variability of pollutants distribution not included; use of arithmetical mean to derive average concentration of PM_{10} values;

Model 3 (Bayesian, Second Version): WinBugs, second version - same variability of Models 1 and 2 plus variability of the pollutants distribution; use of arithmetical mean;

Model 4 (Bayesian, Third Version): WinBugs, third version - same variability of Model 3; use of geometrical mean to derive average concentration of PM10 values; first 1 000 discarded for the calculation of summary statistics.

Model 5 (a, b and c) (Bayesian, Forth Version): WinBugs, forth version - updating of model 4 with historical mortality rates (not estimated ones); same variability of model 3 but data disaggregated by sex and age classes; Two chains of simulated values calculated; 10 000 iterations for each chain, first 1 000 discarded for the calculation of summary statistics; Gelman-Rubin convergence test and convergence methods applied.

Model 6 (Bayesian, Fifth Version): WinBugs, same variability of model 5, calculation extended to years of life lost as well (results not included in the table below).

The results obtained using the above-mentioned different approaches have been shown in the table reported below, where the number of attributable cases estimated by the different models, using three different counterfactual exposure levels (i.e. alternative exposure distribution used as baseline for estimating the burden of disease caused by the exposure distribution of interest) for PM_{10} concentrations: 20, 30 and 40 $\mu g/m^3$ has been reported.

It can be seen that central estimates are not substantially affected by the choice of the model also when more in-depth analyses have been introduced. The substantial differences on the average effects depend on more updated mortality and population data used to calculate the estimates and on the more correct use (on the methodological side) of the geometrical mean as a synthetic indicator. For example, using 20 $\mu g/m^3$ as a counterfactual exposure (first part of the table), the last developed model (5c) estimates that around 6 000 extra deaths are attributable to PM_{10} in the eight major Italian cities each year. While this kind of information is valuable to describe the health benefits associated with abatement measures (and can therefore help formulate evidence-based policies), it is also crucial to be able to evaluate the degree of uncertainty that surrounds the estimates. The tables show the 95% credibility intervals attached to the estimates. Unlike the central estimates themselves, credibility limits are heavily affected by the choice of the model: in particular, models that allow for more sources of variability produce wider credibility intervals (from 2 616 in model 3 to 8 517 in model 5c).

Disaggregating the same results by sex, age-class (single classes or grouped in two classes, working people (between 30 (14, if the health effect under study is acute mortality) and 65 years old) and not working one (more than 65 years old)) and, further, in terms of YLL by age-groups, allow us to compare results between the two groups and among the eight cities and can help the administrators to identify priorities in health impact assessment decision making process.

Table 1. Comparison of results from a deterministic to a probabilistic approach

Baseline 20 - sum of the 8 cities	**mean**	**CI 95% (LL)**	**CI 95% (UL)**	**2.5 perc**	**median**	**97.5 perc**	**CI width**	**N° of Simul.**	**Notes**	**Type of mean used**
1 – No Bugs: var of RR (correct), population > 30	5143	3890	6396	-	-	-	2505	0	Variability of RR	Arithmetical
2 - Bugs: var of RR, population > 30	5105	-	-	3768	5120	6384	2616	10000	Variability of RR	Arithmetical
3 - Bugs: var of RR and C, population > 30	4812	-	-	1110	4787	8642	7530	10000	Variability of RR and C	Arithmetical
4 - Bugs: var of RR and C, population > 30	5018	-	-	1909	4752	9708	7799	10000*	Variability of RR and C	Geometrical
5a - Bugs: var of RR and C, men, one year age-classes	3025	-	-	1227	2894	5649	4422	10000**	Variability of RR and C	Geometrical
5b - Bugs: var of RR and C, women, one year age-classes	2979	-	-	1084	2854	5426	4342	10000**	Variability of RR and C	Geometrical
5c - Bugs: var of RR and C, total, one year age-classes	6003	-	-	2463	5746	10980	8517	10000**	Variability of RR and C	Geometrical

Baseline 30 - sum of the 8 cities	**mean**	**CI 95% (LL)**	**CI 95% (UL)**	**2.5 perc**	**median**	**97.5 perc**	**CI width**	**N° of Simul.**	**Notes**	**Type of mean used**
1 – No Bugs: var of RR (correct), population > 30	3496	2598	4394	-	-	-	1797	0	Variability of RR	Arithmetical
2 - Bugs: var of RR, population > 30	3479	-	-	2540	3488	4391	1851	10000	Variability of RR	Arithmetical
3 - Bugs: var of RR and C, population > 30	3158	-	-	-745	3151	7026	7771	10000	Variability of RR and C	Arithmetical
4 - Bugs: var of RR and C, population > 30	3376	-	-	2046	3107	8198	6152	10000*	Variability of RR and C	Geometrical
5a - Bugs: var of RR and C, men, one year age-classes	2040	-	-	190	1905	4620	4430	10000*	Variability of RR and C	Geometrical
5b - Bugs: var of RR and C, women, one year age-classes	2005	-	-	198	1880	4512	4314	10000*	Variability of RR and C	Geometrical
5c - Bugs: var of RR and C, total, one year age-classes	4045	-	-	386	3781	9155	8769	10000**	Variability of RR and C	Geometrical

Baseline 40 - sum of the 8 cities	**mean**	**CI 95% (LL)**	**CI 95% (UL)**	**2.5 perc**	**median**	**97.5 perc**	**CI width**	**N° of Simul.**	**Notes**	**Type of mean used**
1 – No Bugs: var of RR (correct), population > 30	1765	1269	2262	-	-	-	992	0	Variability of RR	Arithmetical
2 - Bugs: var of RR, population > 30	1764	-	-	1248	1767	2257	1009	10000	Variability of RR	Arithmetical
3 - Bugs: var of RR and C, population > 30	2051	-	-	-2779	1453	5340	8119	10000	Variability of RR and C	Arithmetical
4 - Bugs: var of RR and C, population > 30	1642	-	-	-1800	1387	6552	8332	10000*	Variability of RR and C	Geometrical
5a - Bugs: var of RR and C, men, one year age-classes	1000	-		-963	878	3671	4634	10000**	Variability of RR and C	Geometrical
5b - Bugs: var of RR and C, women, one year age-classes	976	-	-	-935	866	3566	4501	10000**	Variability of RR and C	Geometrical
5c - Bugs: var of RR and C, total, one year age-classes	1977	-	-	-1896	1746	7245	9141	10000**	Variability of RR and C	Geometrical

*In the forth model developed with WinBugs, 10 000 simulations have been made but the first 1 000 have not been used in the analysis.

**In the fifth model, 10 000 simulations have been made for both chains but the first 1 000 have not been used in the analysis.

2.6 The role of uncertainty

There are several uncertainties and limitations deriving from the use of epidemiological evidence to predict health effects as a result of change in ambient pollution concentrations. In the last years some of these uncertainties have been solved as new research was carried out. Some of these uncertainties are of theoretical order, some other practical ones and imply possible light changes to the generic WinBugs model illustrated in § 2.4.

Among the theoretical sources of uncertainties, the issue of causality has to be stressed. Epidemiological studies can show a positive relationship between outdoor air pollution exposure and adverse health effects but they cannot prove the causality of the association. In recent studies, epidemiological evidence has been strengthened by hundreds of new studies carried out worldwide, both on chronic and acute effects, which found a similar magnitude of effect for mortality and have been linked to a wide spectrum of morbidity outcomes. In addition to that, toxicological evidence has been corroborated, even though not all the underlying biological mechanisms have been fully clarified. For these reasons, more than in the past, adverse health effects can be causally linked to changes in ambient air pollution.

However many uncertainties are still present and must be underlined.

A first issue regards the choice of the risk coefficients applied to health impact assessment studies (the complete list will be described in D11.3). They are usually derived from epidemiological studies carried out in other locations and then transferred to other regions of the world. With regard to chronic effects, the few U.S. cohort studies that are applied have been reanalyzed and their robustness has been internationally accepted. With regard to acute effects, time-series studies have been replicated in many cities worldwide and the magnitude of the effects is consistent. On this basis, it does not seem that the transferability and application of these mortality risk coefficients is unreasonable; the width of confidence interval, applied inside the WinBugs model, can reflect part of this uncertainty.

With regard to morbidity effects, ranging from hospital admissions to loss of working days, uncertainty is larger. Health effects are registered in a different way across the world and local situations can differ substantially. In addition to that, another source of uncertainty is the frequent unavailability of reliable background rates. Several of this information can be derived from external situation to estimate impact functions (Hurley, F. et al., 2005), confidence interval are wider and, several times, risk estimates are not statistically significant.

A second issue regards (i) the existence of a threshold below which no adverse health effects can be observed and (ii) the linear approximation of the concentration response functions. The most recent scientific evidence cannot affirm the existence of a threshold and effects should be calculated until background concentrations both for chronic (World Health Organization, 2003) and acute effects (Daniels et al., 2000, Dominici et al., 2003, Pope, 2000, Samoli et al., 2005, Schwartz et al., 2001, Schwartz & Zanobetti, 2000). A linear concentration-response relationship has been supported by several studies (Cohen et al., 2004, Ostro, 2004): only at higher concentrations, rarely observed in European cities, a log-linear function should be more plausible. The use of a linear risk function provides a good approximation of the effects.

A third issue is given by possible errors and limitations in measuring exposure. Fixed-site monitoring stations are generally used for the calculation of an average concentration value, which should give a good approximation of the community exposure. Individual exposure and personal time activity patterns and are not taken into account. However, health impact assessment studies are generally not sophisticated on personal exposure side and risks are calculated on large populations. Outdoor air pollution is a ubiquitous risk and an average value, together with internationally accepted criteria used to choose a representative subset of the monitoring stations, should provide an average value representative for the exposure of the general population. In addition to that, the extrapolation of risks calculated on the general population to this study does not allow the estimates of health impact for people belonging to more sensitive groups, for instance, the poorest socio-economic groups in the population. This last aspect will be treated in deliverable D11.3.

A forth issue derives from co-pollutants: only PM_{10} has been considered because of its high correlation with other urban pollutants. Double counting is avoided but an underestimation of the health effects is certain.

On the methodological side, with practical implications for the Bayesian model illustrated before, additional source of uncertainties derive from the obligatory use of conversion and correction coefficients used to have homogeneous values (i.e. gravimetric concentrations in $PM_{2.5}$ scale) on which apply health impact assessment methods. This kind of uncertainty cannot be solved before a substantial improvement and modernization of the national network of monitoring stations. Therefore, some methods have been proposed (APHEIS (Air Pollution and health: A European Information System), 2005) and are illustrated below.

In fact, three different methods are used to measure PM_{10} concentrations: BETA automatic, TEOM (Tapered Element Oscillating Microbalance) and gravimetric. The method recommended by the European legislation is the gravimetric. It has been demonstrated that the use of TEOM and beta methods underestimates PM_{10} above all at high level of concentrations (Stanger Science & Environment, 1999) and a correction coefficient of 1.3 to annual averages is needed (EC working group on particulate matter, 2004) in order to compensate the losses of volatile particulate matter. In addition to that, the risk estimates calculated in U.S. cohort studies proposed in this report (D11.3) derive from monitoring stations using gravimetric methods. For these reasons, in this study the following correction coefficients are proposed:

$$Beta\ Automatic\ PM_{10} * 1.3 = Gravimetric\ PM_{10}$$

and

$$TEOM\ \ PM_{10} * 1.3 = Gravimetric\ PM_{10}$$

Sometimes local conversion factors can be available and reliable. In this case some scaling factor inside the WinBugs model, according to the city-specific situation, can be introduced.

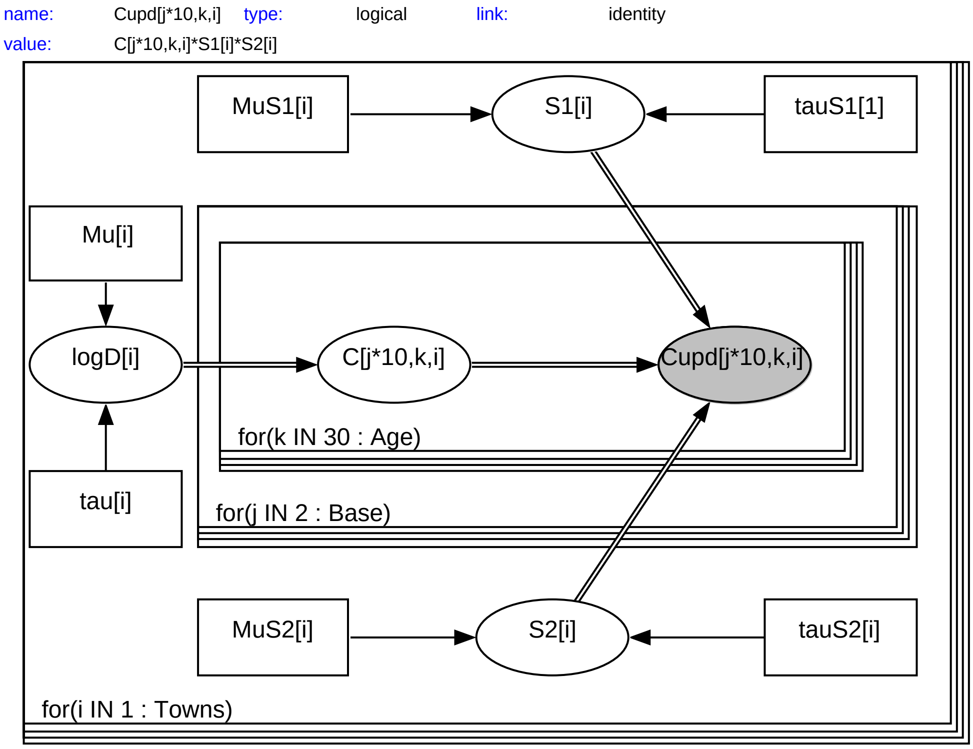
Another methodological issue regards the conversion of PM_{10} values concentration to $PM_{2.5}$ in the case in which relative risks are available only in terms of $PM_{2.5}$ metric. $PM_{2.5}$ is not routinely measured yet in several European countries and a conversion factor between PM_{10} data and $PM_{2.5}$ is needed to use those risk estimates. Annual mean $PM_{2.5}$ levels are roughly two thirds those of PM_{10} with substantial variations in time and space (ratios from 0.4 to 0.8 have been reported (CAFE working group on Particular Matter, 2004). In a recent volume by WHO on comparative quantification of health risk written by the “Global Burden of Disease” (GBD) group(Ezzati et al., 2005), urban air pollution has been considered as one of the major risk factors. In this study the basic ratio between $PM_{2.5}$ and PM_{10} has been assumed at 0.5 even though, as the authors claim, higher and lower levels have been observed (Cohen et al., 2004). If no reliable local conversion factors were available a default factor of 0.7 should be used, as recommended by the “Air Pollution and Health: A European Information System” (APHEIS) study (APHEIS (Air Pollution and health: A European Information System), 2005). This estimate is based on the average, weighted with the inverse of the standard errors, coming from two recent studies. In the first one (CAFE working group on Particular Matter, 2004), the second position paper on the particulate matter by the CAFE group, based on 72 European locations, a ratio of $PM_{2.5}/PM_{10} = 0.65$ (standard error = 0.09) has been found; in the second one, based on 11 stations, a ratio of $PM_{2.5}/PM_{10} = 0.73$ (standard error = 0.15) (Van Dingenen et al., 2004) has been reported.

For this reason the following conversion coefficient should be applied:

$$PM_{2.5} = 0.7 * PM_{10}$$

However, as with measurement methods, sometimes local precise conversion factors can be available. Also in this case, according to the city-specific situation, inside the WinBugs model some scaling factors could be introduced.

In the Bayesian model the practical implication should be that the “C_{upd}” ellipse, representing the updated “change” in concentration should be obtained from the original “C” ellipse multiplied by two scaling factors, as illustrated below:



name: Cupd[j*10,k,i] type: logical link: identity

value: C[j*10,k,i]*S1[i]*S2[i]

corresponding to the following WinBugs script

```
for( i in 1 : Towns ) {
    for( k in 30 : Age ) {
      for( j in 2 : Base ) {
        C[j * 10 , k , i] <- exp(logD[i]) - (10 * j)
                 Cupd[j * 10 , k , i] <- C[j * 10 , k , i] * S1[i] * S2[i]
      }     }   }
  for( i in 1 : Towns ) {
    S1[i] ~ dnorm(MuS1[i],tauS1[1])
    S2[i] ~ dnorm(MuS2[i],tauS2[i])
    logD[i] ~ dnorm(Mu[i],tau[i])
```

where:

- C, logD, Mu and tao have been already described in the first model;

- S1 is the first scaling factor (correction from TEOM and/or Beta to gravimetric) depending on Mu[S1] and tau[S1], formed by 8 element vectors. For instance Mu[S1] can be equal to (1.3, 1.3, 1.15, 1.1, 1.3, 1.45, 1, 1.1) where 1.3 is the "standard" correction coefficient described before, 1 corresponds to a city for which a correction is not needed and the other ones are values locally estimated. Tau[S1] represents further variability attached to the local estimates which can be included in the model;

- S2, to be introduced in the model if the risks are calculated in $PM_{2.5}$ metric, , is the second scaling factor and represents the $PM_{2.5}/PM_{10}$ conversion. It depends on two 8 element vectors, Mu[S2]and tau[S2]. For instance Mu[S2] can be equal to (0.7, 1, 0.7, 0.55, 1, 0.7, 0.65, 0.9) where 0.7 is the standard conversion coefficient described before, 1 corresponds to a city in which $PM_{2.5}$ data are available and the other ones are conversion coefficients locally estimated (each of them with further variability attached, described by tau[S2]);

- C_{upd} is the updated C stochastic node and is equal to C*S1*S2

As anticipated, the use of these coefficients further enlarges the width of the credibility intervals and gives a more exact idea of the precision of the estimates.

3. A BAYESIAN EXPOSURE MODEL

3.1 Introduction

Pollution exposure models in small area analyses are commonly based solely on the outdoor pollution level at the place of residence (Best et al., 2000), Maheswaran et al. (2005). When using routinely collected data, place of residence is often the only piece of information available by which to fix a subject's (and by extension the area's) exposure level. Here, we aim to explore how we might combine data from two routine surveys in order to obtain additional estimates of exposure levels associated with place of work and transport.

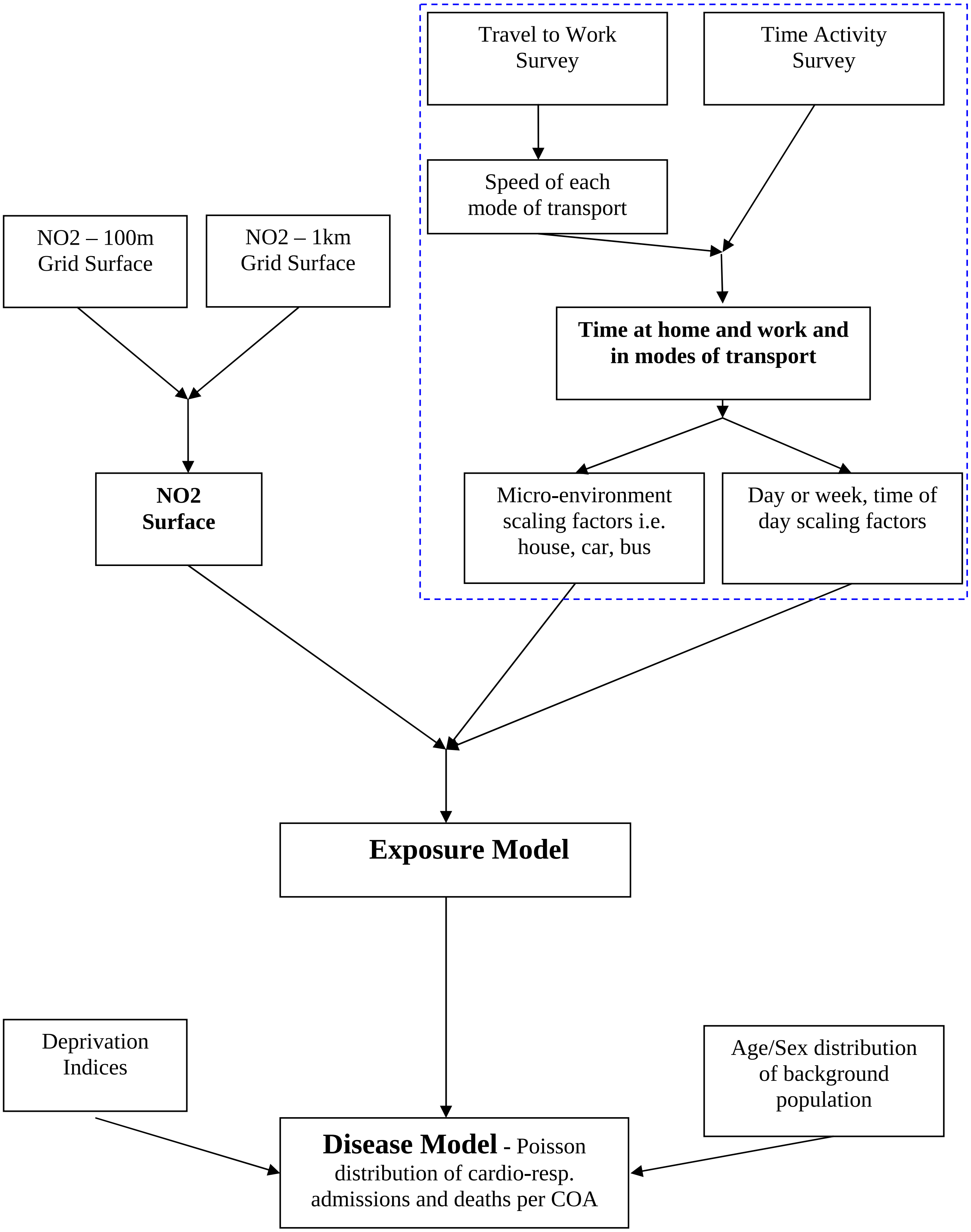
Previous work in probabilistic exposure models includes Zidek et al. (2005), Lee et al. (2000) and Ozkaynak et al. (1995) in modelling exposure levels in microenvironments. These models incorporate probabilistic considerations such as the number of smokers in the household, whether the stove is gas or electric and meteorological data. This detail is too ambitious for the type of routine survey data (see § 3.2 below) we wish to make use of. A further reason for limiting the complexity of our models is that we seek to make inference for broad sections of the populations defined only by age, sex and postcode of residence. Zidek et al. by comparison, who consider predictions for example for "women who live in semi-detached Brent dwellings, work in Bloomsbury, smoke, and cook with gas", study highly specified populations.

In addition to the place of residence we set out to provide information on the distribution of possible work places and some information on the time spent in various modes of transport. We make use of data from the area of the city of Sheffield and its immediate surroundings. From an OPUS point of view it might have been preferable to carry out the modelling in the London area. However, although the proposed methodology is general and in principle applicable to any region of interest or pollutant of choice, the relative homogeneity in the background NO_2 levels of London and the widespread use of the underground together with the apparent lack of data on pollution levels in this environment mean that London is not an ideal choice of region.

A representation of the overall model, encompassing the exposure, pollution surface and disease sub-models, is given in Figure 1. The area enclosed by the dashed line is the activity model described below which combines with an NO_2 pollution surface in order to inform our exposure model.

The remainder of this section of the deliverable is presented as follows: in § 3.2 we describe the survey data sets we use in this part of the model; in § 3.3 we set out the model used to combine these data sets and give the associated graphical model; in § 3.4 we list the WinBugs code used to implement this model and in § 3.5 we consider the work which follows on from this modelling strategy.

Figure 1 – Model Overview



3.2 Data Sets

3.2.1 Transport to Work Survey

The Origin-Destination statistics

(http://www.statistics.gov.uk/STATBASE/Product.asp?vlnk=13900&More=Y) derived from the 2001 Census cover flows at various different geographical levels for migration and travel to work. In particular we make use of the travel to work data at the Output Area (OA) level. Output Areas typically contain around 125 households and the 175,000 Output Areas of the UK 'nest' within wards and parishes, and normally comprise of whole unit postcodes (See http://www.statistics.gov.uk/census2001/). All people aged 16-74 and in employment are included. However, small counts are adjusted to prevent the disclosure of information about identifiable individuals.

Each line in the files relates to a single combination of OA of origin (home) and OA of destination (usual place of work). The number of people using each mode of transport as the main means of their travelling to work are recorded for each OA to OA combination. The groupings of mode of transport are as follows:

- Underground, metro, light rail, tram
- Train
- Bus, mini bus, coach
- Motorcycle, scooter, moped
- Driving a car or van
- Passenger in a car of van
- Taxi or minicab
- Bicycle
- On foot
- Other

A column is also included for those people who work mainly from home. This column is by definition only populated where the origin and destination OA are the same.

A geographical link is made by adding the population weighted centroid of each OA to the dataset. We are thus able to place each OA, either of residence or workplace, in a grid square of the NO_2 surface.

A Fortran program has been written which describes the straight line path from home OA to work OA. The program gives an exposure measure for each route based on the relative distances travelled in each of the NO_2 grid squares along the route. Assuming a direct route between the OA’s is simplistic but does have the computational advantage that sophisticated routing choices over road and rail networks are not required. A future addition to this work might include transport network based routing.

When computing the time spent in each mode of transport to work it must be noted that the distance between home OA and work OA is the total distance travelled in all modes of transport to work, not the distance travelled while in the main mode of transport to work. This survey only reports the main mode of transport used on this journey, it does not include secondary modes of transport. We wish to specifically infer the time spent in the main mode of transport and so we must make use of the Time Use Survey in order to give a split between main and secondary modes of transport.

3.2.2 UK 2000 Time Use Survey

The main aim of the UK 2000 Time Use Survey (http://www.statistics.gov.uk/timeuse/) was to measure the amount of time spent by the UK population on various activities. The survey comprised private households and the household members living in those households.

Selected household heads or their partners completed a household questionnaire. All individuals aged 8 or over were asked to complete individual questionnaires, two 24-hour diaries and a one week work and education time sheet.

The household and individual questionnaires were mainly used to gather background information and demographics. The diaries are broken down into ten minute slots for which primary and secondary activities as well as information on the respondent's location are recorded.

From 11 664 individuals there were a total of 20 981 diary responses, of which 19 898 were deemed by ONS to be of the standard required for subsequent analysis. We further reduce the number of diary responses to 15 229 by excluding those aged under 16 and those in the London and Northern Ireland regions. These geographically based exclusions were as a result of the transport use profile for these two regions appearing to be distinct from the rest of the country.

We concentrate on demographic variables and the duration of spells in specific locations. In particular, we investigate the time spent during each diary day in various transport modes. This was further split into time spent travelling to or from work and time spent on non work travel.

The raw transport modes were similar to those listed above and for the purposes of analysis further aggregation of the modes of transport was necessary. This was in a desire to simplify but it was carried out with pollution exposure in mind. For instance we might assume that the level of exposure to NO_2 would be similar for car drivers and car passengers and further that levels within cars and taxis would be similar. It is also important when grouping modes that the relative speeds of each mode should be taken into account. For instance, should motorcyclists be grouped with bicyclists or with car users? The NO_2 level experienced by motorcyclists might be expected to be somewhere in between the enclosed environment of a car travelling in the centre of the traffic and the open environment of a bicyclist which travels slightly to the side of the main traffic flow. However, a second consideration is the relative speeds at which the bicyclist, motorcyclist and car travel. Motorcyclists travel at a similar speed to car users, enduring a similar duration within their pollution environment; thus it was decided to group motorcyclists and car users together. The full groupings are as follows:

Group 1: Walking

Group 2: Bike

Group 3: Train, Tram, Tube

Group 4: Car (driving/passenger), Lorry/Tractor, Van, Taxi, Moped, Motorcycle

Group 5: Bus, Coach

Those diary slots recording “other mode of transport” were not included since it would not be possible to ascertain the pollution environment associated with such a catch-all miscellaneous code.

3.2.3 Additional data sets

A number of other data sets which will not be described in detail in this deliverable, will also be used in the final model. Hospital Episode Statistics (HES) data, which is routinely collected each year, will provide age/sex/postcode identified health outcomes. Two pollution grids, available at differing resolutions, will be combined to provide the underlying pollution surface. Finally, various scaling factors and constants, such as the speed of various modes of transport (see Transport Statistics Bulletin reference) and relative pollution levels of home and workplace, will have priors obtained from the literature assigned to them.

3.3 A Combined transport use model

We now describe the model which will be used to integrate the Transport to Work and Time Use surveys and in § 3.2.2 we give the model in graphical form.

In order to link the two surveys we must infer the main mode of transport used to travel to work in the diary respondents. The time spent travelling to work during each 24 hour diary response is totalled for each mode of transport grouping. The main mode of transport to work for each diary entry is simply defined as that mode in which most time was spent.

The variables we model are the sum, for each of the 5 transport modes, of non travel-to-work time and travel-to-work time via modes other than the main mode of transport to work. For example, for a diary day summarised as follows:

Travel-to-work: Walk 10 mins, Bike 0 mins, Train 30 mins, Car 0 mins, Bus 0 mins

Non travel-to-work: Walk 20 mins, Bike 0 mins, Train 10 mins, Car 40 mins, Bus 0 mins

The main mode of transport to work in this example would be Train, and the combined response which we seek to model could be termed the Non Main-Work-Mode (NMWM) travel time and would for this example be:

NMWM Walk 30 mins, Bike 0 mins, Train 10 mins, Car 40 mins, Bus 0 mins

Thus in this example the 30 minutes spent travelling to work by train is excluded from the NMWM total since this is the main mode of transport to work. All other time spent travelling to work, in this example the 10 minutes walking, is included in the NMWM total.

We are able to arrive at the total time spent travelling during a 24 hour period by summing the time spent in the main mode of transport-to-work (taken from the Travel to Work survey) and the modelled response, NMWM, from the Time Use survey.

We choose not to model the Travel to Work Survey but instead to use its data as a probabilistic model. There are rounding issues in the survey data, caused by the changes necessary to prevent identifiability of individuals, but we do not include this extra complexity here.

The main feature of this NMWM data which we wish to capture is the interaction and correlation between people's choice of transport mode. For instance, those who have taken a train to work are highly likely to have undertaken some walking also, either to or from the station. Similarly those who cycle to work might be expected to be less likely to walk since most distances covered on foot would be just as easily covered by a cyclist.

3.3.1 Statistical Model

For each mode of transport we split the NMWM data, which is already in the form of discrete blocks of ten minutes resulting from the dairy format, into ordered categorical data using 4 bins :- 0 minutes, 10-30 mins, 40-60 mins and 70+mins. For ease of writing let us denote our response NMWM as T.

Formally we can write that we have j = 1,…5 modes of transport for each i = 1,…n diary responses and that each of these can take one of k = 1,…4 categorical values.

We therefore wish to model

$$\Pr\{T_{ij} = k \mid X_i\} = p_{jk} \text{ for } i = 1,\ldots n,\ j = 1,\ldots 5, k = 1,\ldots 4$$

and covariate information X_i, subject to the constraints $0 < p_{jk} < 1$, for $j = 1,\ldots 5, k = 1,\ldots 4$ and also $\sum_{k=1}^{4} p_{jk} = 1$ for $j = 1,\ldots 5$.

As prior distributions for each of the p_{jk} we use uniform distributions on the interval (0, 1).

For simplicity of explanation let us assume that X_i is univariate, and contains categorical covariate information. In building the model further we include this covariate information X_i via p_{j1}, the probability that the realisation of mode j takes value k=1 i.e. 0 mins.

So the unscaled "probabilities" for each of the k = 1,…4 categorical values are now also subscripted by i, and are given by

$$p'_{ij1} = p_{j1} \times \beta_{js} \text{ with } p'_{ijk} = p_{jk} \text{ for } k = 2,3,4$$

where

$$\beta_{js} = \begin{cases} \beta_{j1} & \text{if } X_\mathrm{i} = \text{ the reference category} \\ \beta_{j2} & \text{if } X_i \text{ otherwise} \end{cases} . [1]$$

And in order to maintain the constraint of summing to unity the p_{ijk} are rescaled as follows

$$p_{ijk} = p'_{ijk} / \sum_{k=1}^{4} p'_{ijk} \text{ for } k = 1...4 \ .$$

All covariates enter the model for p_{ij1} multiplicatively thus maintaining positivity. Identifiability is ensured by setting the first level of the covariate effect, β_{j1}, to the null effect value for multiplicative models of 1.0. For those β_{js} not corresponding to the reference category and hence set to 1.0, we assign non-informative prior distributions.

In this formulation the covariate information affects the total weight of probability assigned to non-zero values, $\sum_{k=2}^{4} p_{ijk}$, but does not alter the distribution of the probability between the p_{ij2}, p_{ij3} and p_{ij4}.

The covariates considered for the model were

- main mode of transport to work, corresponding to groups 1-5
- age of the respondent, split into 3 bands
- sex of the respondent, M or F
- population density of the area of the respondent's household, split into 3 bands
- diary day, weekday or weekend
- household income, split into 5 bands.

We also included, for j = 1,…5, a 4-dimensional vector of covariates, α_j. Each vector contains elements $\alpha_{jj'}$, defined for the four values of j' not equal to j.

Each element $\alpha_{jj'}$ is defined as follows

$$\alpha_{jj'} = \begin{cases} \alpha_{jj'1} & \text{if } T_{\text{ij'}} \text{ in category } 1 \\ \alpha_{jj'2} & \text{if } T_{ij'} \text{ in categories 2,3 or 4} \end{cases}$$

Thus where $\text{T}_{\text{ij'}}$ is in category 1, corresponding to zero minutes travelled in mode j', the term $\alpha_{jj'1}$ is set to the null effect value of 1.0 but where $\text{T}_{\text{ij'}}$ is in categories 2, 3 or 4, corresponding to a at least one ten minute diary slot being allocated to transport mode j', then the estimated value of $\alpha_{jj'2}$ enters the model.

The α_j enter the model multiplicatively, in the same way as the β parameters and are assigned similar prior distributions. This α_j covariate provides a means of interaction between the different modes of transport in addition to that supplied by the main mode of transport to work covariate.

The model is implemented and evaluated using Bayesian Markov chain Monte Carlo (MCMC) methods via WinBugs 1.2 (Spiegelhalter et al., 1999).

3.3.2 GAPM and Graphical Model

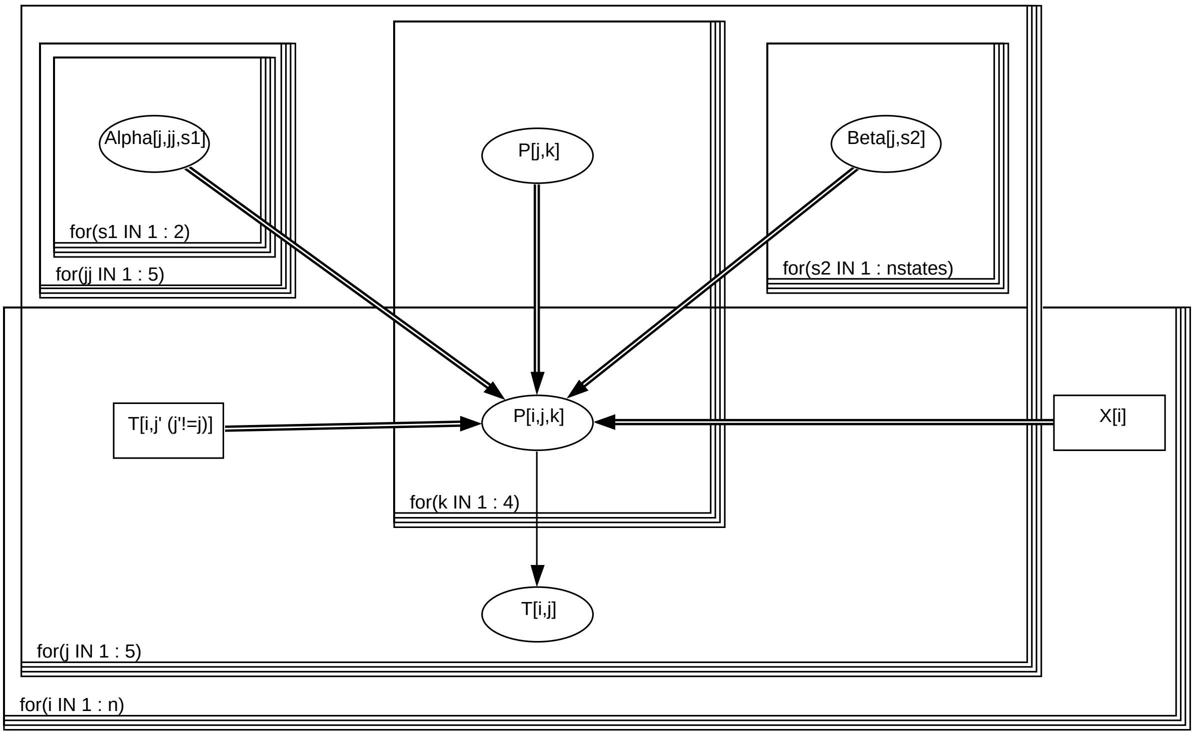
A graphical a priori model (GAPM) as described in WP06, is currently being generated for this model.

A traditional graphical model representation of the model is given in Figure 2. Whilst this was drawn using DoodleBUGS (the graphical model interface of WinBUGS), this is a simplification of the DoodleBUGS model which would be needed to recreate the WinBUGS code as written in § 3.4.

The graph is shown only including $\alpha[]$ and $\beta[]$ as defined above. Further covariate influences would be shown in a similar way to that shown for $\beta[]$.

As described in § 3.3.1, the X[i] and T[i,j' (j'!=j)] data boxes directly affect which elements of respectively $\beta[]$ and $\alpha[]$ are used in the model for T[i,j]. The T[i,j' (j'!=j)] therefore corresponds to the four modes of transport other than mode j, the response variable which is currently being modelled. For simplicity the $P'[i,j,k]$ from [1] above, which would have appeared as an intermediate calculation step in between $P[j,k]$ and $P[i,j,k]$, are omitted from the graph and instead $\alpha[]$ and $\beta[]$ are drawn influencing $P[i,j,k]$ directly.

Figure 2 – Graphical Model for Transport Time Use



Where :-

i denotes diary day

j denotes the mode of transport

k denotes the four categories of possible response

jj loops through the transport modes

and s1 and s2 denote the states of the covariate effects.

3.4 WinBugs model code

Listed below is the code used to run the model. Initial estimates are also shown. The inclusion of covariates is as for the graphical model. P3[j,k] corresponds to the p_{jk} in [1] but P2[j,k] and P1[ ] are merely used as intermediate calculation steps. And s1_1,…,s1_5 correspond to the index s1 for values of the loop jj = 1,…5.

```
model{
  for(i in 1:nrows){

    for(j in 1:5){
        IT[i,j] <- step(T[i,j]-1.1)+1    #Indicator Variables used to  identify appropriate P
    }

    for(j in 1:5){
        T[i,j] ~ dcat(P[IT[i,1],IT[i,2],IT[i,3],IT[i,4],IT[i,5],j,MM[i],1:4])  # where MM[i] is
                                                                      # the main mode of
                                                                           # transport to work
    }
  }
##### Code for P[ ] vector

  for(j in 1:5){                # Prior on P
    for(k in 1:4){
        P3[j,k] ~ dunif(0.0,1.0)
    }
  }

  for(j in 1:5){                # Enforces sum to unity constraint
    for(k in 1:4){
        P2[j,k] <- P3[j,k]/sum(P3[j,1:4])
    }
  }

  for(s1_1 in 1:2){             # Vector of P values dependent on covariate values
    for(s1_2 in 1:2){
      for(s1_3 in 1:2){
        for(s1_4 in 1:2){
          for(s1_5 in 1:2){
            for(j in 1:5){
              for(s2 in 1:6){   # Loop over modes of Main transport to work

                P1[s1_1,s1_2,s1_3,s1_4,s1_5,j,s2,1] <-
        P2[j,1]*A[j,1,s1_1]*A[j,2,s1_2]*A[j,3,s1_3]*A[j,4,s1_4]*A[j,5,s1_5]*B[j,s2]
                P1[s1_1,s1_2,s1_3,s1_4,s1_5,j,s2,2] <- P2[j,2]
                P1[s1_1,s1_2,s1_3,s1_4,s1_5,j,s2,3] <- P2[j,3]
                P1[s1_1,s1_2,s1_3,s1_4,s1_5,j,s2,4] <- P2[j,4]


              for(k in 1:4){
```

```
                    P[s1_1,s1_2,s1_3,s1_4,s1_5,j,s2,k] <-
  P1[s1_1,s1_2,s1_3,s1_4,s1_5,j,s2,k]/sum(P1[s1_1,s1_2,s1_3,s1_4,s1_5,j,s2,1:4])
                  }
                 }
                }
               }
             }
           }
         }
       }

##### Prior Distributions for covariate parameters

  for(j in 1:4){          # A corresponds to alpha in Fig 2, mode of transport interaction
    for(jj in (j+1):5){
         A[j,jj,1] <- 1.0
         A[j,jj,2] ~ dunif(0.0,2.0)
         A[jj,j,1] <- A[j,jj,1]
         A[jj,j,2] <- A[j,jj,2]
    }
  }

  for(j in 1:5){
    A[j,j,1] <- 1.0
    A[j,j,2] <- 1.0

    for(k in 1:5){        # B corresponds to Beta in Fig 2, main mode of work transport
         B[j,k] ~ dunif(0.0,2.0)
    }
    B[j,6] <- 1.0      # For those who work at home or are unemployed
  }

}


Initial Estimates
list(
T3P=structure(.Data = c(
0.5,0.3,0.1,0.1,
0.5,0.3,0.1,0.1,
0.5,0.3,0.1,0.1,
0.5,0.3,0.1,0.1,
0.5,0.3,0.1,0.1), .Dim = c(5,4)),

B=structure(.Data = c(
1.0,1.0,1.0,1.0,1.0
,NA,
1.0,1.0,1.0,1.0,1.0,NA,
1.0,1.0,1.0,1.0,1.0,NA,
1.0,1.0,1.0,1.0,1.0,NA,
1.0,1.0,1.0,1.0,1.0,NA), .Dim = c(5,6)),
```

```
A=structure(.Data = c(
NA,NA,NA,  1,NA,  1,NA,  1,NA,  1,
NA,NA,NA,NA,NA,  1,NA,  1,NA,  1,
NA,NA,NA,NA,NA,NA,NA,  1,NA,  1,
NA,NA,NA,NA,NA,NA,NA,NA,NA,  1,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA ), .Dim = c(5,5,2)) )
```

3.5 Future work

The framework for the overall model which makes use of these pieces of exposure information must be decided upon. A choice has to be made between a small area study or a case control approach. It may turn out that the latter better suits the manner in which we make predictions from the models described above.

The way that these pieces of exposure information are introduced into the disease model is also not obvious. Further decisions about whether the separate parts of the activity prediction, namely the home, work and transport elements should be entered independently into the disease model or whether they should somehow be integrated into one exposure figure using scaling factors which relate the background pollution level to the level in the house/office/mode of transport.

The simplistic nature of the routing between places of work and home (see § 3.2.1 above) may well argue in favour of including the transport element as a separate exposure risk. Further, the uncertainty associated with each of the three elements varies greatly, from reliable data originating from the HES home address field for place of residence, to modelled values which themselves rely on a series of modelling assumptions and inherent uncertainty.

Different modelling strategies will be explored in order to take account of these factors and to suggest the best way to make use of the separate exposure predictions.

4. REFERENCES

APHEIS (Air Pollution and health: A European Information System) (2005). *Health Impact Assessment of Air Pollution and Communication Strategy. Third-year Report 2002-2003 (APHEIS-3).* Institute de Veille Sanitaire,European Commission, Agència de salut publica, Saint-Maurice, France.

Best NG et al. (2000). Combining models of health and exposure data: the SAVIAH study. In: Elliott P et al., eds. *Spatial epidemiology. Methods and applications*. Oxford: Oxford University Press; p. 393-414.

Brooks SP, Gelman A (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics,* 7 434-455.

CAFE working group on Particular Matter (2004). *Second Position Paper on Particular Matter - Final Draft*.

Cohen JC et al. (2004). Urban air pollution. In: Ezzati M et al., eds. *Comparative Quantification of Health Risks. Global and regional burden of disease attributable to selected major risk factors. Volume 2*. Geneva: World Health Organization; p. 1353-1433.

Daniels MJ et al. (2000). Estimating particulate matter-mortality dose-response curves and threshold levels: an analysis of daily time-series for the 20 largest US cities. *Am J Epidemiol,* 152 (5):397-406.

Dominici F et al. (2003). Shape of the exposure-response relation and mortality displacement in the NMMAPS database. In: *Revised analyses of time-series studies of air pollution and health*. Boston: Health Effects Institute; p. 91-96.

EC working group on particulate matter (2004). *Guidance to the member states on PM10 monitoring and intercomparisons with the reference method*.

Ezzati M et al., eds (2005). *Comparative Quantification of Health Risks. Global and regional burden of disease attributable to selected major risk factors. Volume 2*. Geneva.

Gelfand AE, Hills SE, Racine-Poon A (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J Am Stat Assoc,* 85 972-985.

Gelman A, Rubin DB (1992). Inference from iterative simulations using multiple sequences. *Statistical Science,* 7 457-511.

Hurley F et al. (2005). *Methodology for the Cost-Benefit analysis for CAFE: Volume 2: Health Impact Assessment*. (AEAT/ED51014/Methodology Volume 2 Issue 1).

Hurley JF et al. (2000). *Towards assessing and costing the health impacts of ambient particulate air pollution in the UK*. Institute of Occupational Medicine, Edinburgh.

Kunzli N et al. (1999). *Health costs due to road traffic-related air pollution. An impact assessment project of Austria, France and Switzerland: Air pollution attributable cases.* (Prepared for the third Ministerial Conference for Environment and Health, London 1999).

Lee K, Yang W, Bofinger ND (2000). Impact of microenvironmental nitrogen dioxide concentrations on personal exposures in Australia. *J Air Waste Manag Assoc,* 50 (10):1739-1744.

Maheswaran R et al. (2005). Outdoor Air Pollution and Stroke in Sheffield, United Kingdom: A Small-Area Level Geographical Study. *Stroke,* 36 (2):239-243.

Martuzzi M et al. (2002). *Health Impact Assessment of air pollution in the eight major Italian cities*. World Health Organization - European Centre for Environment and Health, (EURO/02/5040650).

Martuzzi M, Krzyzanowski M, Bertollini R (2003). Health impact assessment of air pollution: providing further evidence for public health action. *Eur Respir J Suppl,* 40 86s-91s.

Murray C, Lopez AD, eds (1996). *The Global Burden of Disease*. Geneva: World Health Organization, Harvard School of Public Health, World Bank.

Ostro B (2004). *Outdoor air pollution. Assessing the environmental burden of disease at national and local levels*. World Health Organization, Protection of the Human Environment, Geneva (Environmental Burden of Disease Series, No 5).

Özkaynak H et al. (1995). Errors in Estimating Children's Exposures to NO2 Based on Weeklong Average Indoor NO2 Measurements. In: Morawska L, Bofinger ND, Maroni M, eds. *Indoor Air An Integrated Approach*. Oxford: Elsevier Science; p. 43-46.

Pope CA, III (2000). Invited Commentary: Particulate Matter-Mortality Exposure-Response Relations and Threshold. *Am. J. Epidemiol.,* 152 (5):407-412.

Rabl A (2003). Interpretation of air pollution mortality: number of deaths or years of life lost? *J Air Waste Manag Assoc,* 53 (1):41-50.

Richardson S, Gilks WR (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am J Epidemiol,* 138 (6):430-442.

Roosli M et al. (2005). Years of life lost attributable to air pollution in Switzerland: dynamic exposure-response model. *Int J Epidemiol,* .

Samoli E et al. (2005). Estimating the exposure-response relationships between particulate matter and mortality within the APHEA multicity project. *Environ Health Perspect,* 113 (1):88-95.

Schwartz J et al. (2001). The concentration-response relation between air pollution and daily deaths. *Environ Health Perspect,* 109 (10):1001-1006.

Schwartz J, Zanobetti A (2000). Using meta-smoothing to estimate dose-response trends across multiple studies, with application to air pollution and daily death. *Epidemiology,* 11 (6):666-672.

Spiegelhalter DJ, Thomas A, Best NG (1999). *WinBUGS Version 1.2 User Manual*. Cambridge: MRC Biostatistics Unit.

Stanger Science & Environment (1999). *Assistance with the review and assessment of PM10 in relation to the proposed stage 1 of EU limit values*. (Report prepared for the Department of Environment, Transports and the regions, the Welsh Office and the Scottish Office).

Tierney L (1983). A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *J. Sci. Stat. Comput,* 4 706-711.

Van Dingenen R et al. (2004). A European aerosol phenomenology - 1: physical characteristics of particulate matter at kerbside, urban, rural and background sites in Europe. *Atmospheric Environment,* (38):2561-2577.

World Health Organization (2003). *Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide*. World Health Organization, Bonn (EUR/03/5042688).

Zidek JV et al. (2005). Using a probabilistic model (pCNEM) to estimate personal exposure to air pollution. *Enviromentrics*, 16 481-493.