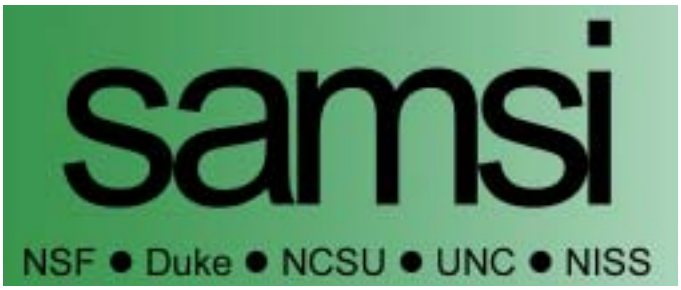

A Computational Model for Estimating Personal Exposure to Air Pollutants with Application to London's PM_{10} in 1997

James V. Zidek, Jean Meloche, Gavin Shaddick,
Chris Chatfield and Rick White

Technical Report #2003-3
March 19, 2003

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

A Computational Model for Estimating Personal Exposure to Air Pollutants with Application to London's PM_{10} in 1997

James V Zidek[1], Jean Meloche[1],
Gavin Shaddick[2], Chris Chatfield[2] and Rick White[1]

[1]University of British Columbia
[2]University of Bath

March 13, 2003

Abstract:

This paper describes a conceptual framework within which models can be developed for predicting the exposure to specified pollutants, of a randomly selected member of a designated population. Such predicted exposures can in turn be used for such things as the analysis of their impacts on human health. The model uses randomly selected time-activity patterns of individual members of the population to determine exposure variations. It can answer questions like: (i) what fraction of the population sustained 'high' levels of exposure, (ii) how many sustained such exposures for say 10 days in a row? (iii) what impact will a proposed new set of air quality criteria have on population members under the age of 4? Over the age of 65?

The framework is developed by abstracting components ('building blocks') of common exposure prediction models for risk assessment along with their linkages, stochastic and structural. Practical considerations in implementing the framework to construct models are considered, including human time-activity patterns and their relationship to environmental factors that help determine such patterns. For example, in summer people may tend to stay indoors in warm weather, while in winter the reverse may be true. Thus, on summer days indoor sources of the hazardous substance may be more important in determining exposures than outdoor sources.

The immediate implementation of the above framework is a computing platform referred to as 'pCNEM' that can be accessed by registered users through the WWW. Such users can then construct a model of the type referred to above, by designating a pollutant of interest, a study area, and a study period. They then incorporate local pollution sources, set source emission and other parameters, and finally upload the requisite data weather as well as pollution data. The paper demonstrates the construction of such a model, one for predicting the exposure to PM_{10} of random selected individuals from sub-populations of Greater London in 1997.

1 Introduction and Summary

This paper presents a methodology for predicting human exposure to an environmental hazard. It addresses the question of the exposure to air pollution particles experienced by individual members (or sub-groups) of the population. More specifically, it considers air pollution consisting of inhalable particles, designated as PM_{10}, that are widely considered to be injurious to human health.

The underlying concepts behind the methodology presented here are well established and have been implemented in a number of models, such as the 'simulation of human activity and pollutant exposure model' (Ott et al 1988), the population-based exposure model of MacIntosh et al (1995), its sequel Burke et al (2001) (hereafter referred to as 'SHEDS') and the forerunner to this work; the 'Probabilistic Version of the NAAQS Exposure Model' (pNEM; Law et al 1997).

Models such as these have played an important role in estimating exposures to air pollution, and the original pNEM model was used in formulating air quality criteria for air pollutants in the United States. The implementation presented here builds on this earlier work, but is original in that instead of a specific model, it provides an underlying WWW platform for developing a wide variety of models. This paper demonstrates the use of that platform and develops a model which can help answer substantive question relating to human exposure to air pollution.

The structure of the paper is as follows. After an introduction, Section 2 describes the general framework for the exposure models of Brauer et al (1996). It describes the components ('building blocks') of models such as SHEDS and pNEM, along with their linkages, both stochastic and structural.

Section 3 considers the practical considerations in the implementation of the framework, including human time-activity patterns and their relationship to environmental factors. This interaction between environment and behavior, though too complex to model, can be represented by means of a catalogue of human behavior patterns. This catalogue contains the patterns revealed in diary data obtained from surveys of human populations. Such a catalogue was obtained from the National Human Activity Pattern Study (NHAPS) a 24 hour recall survey; (Robinson and Thomas, 1991) that was carried out in the United States and three Canadian centers. To simulate human behavior conditional on individual and environmental stratification factors, then requires only that a single item be sampled at random from the catalogue under the assumption that it will serve to represent the population on which the time-activity patterns were surveyed. Such information is required in every application wherein human interaction with the environment is to be represented.

Section 4 presents the implementation of the above framework, a platform referred to as 'pCNEM' (standing ambiguously for a 'probabilistic version of the Canadian version of NEM' and a 'PC version of NEM', the latter referring to the fact that, to the best of the authors' knowledge, it is the first version of NEM to run on a PC). An important feature of this implementation is that it can be accessed and run remotely by (registered) users through the World Wide Web (WWW). Users can then develop a personal exposure model

on that platform and apply it to any specified pollutant, area and time period of interest, using their own data.

Section 5 provides examples of the development and application of the pCNEM model, in order to predict the exposure to PM_{10} of random selected individuals from sub-populations of Greater London in 1997. This begins with an analysis of hourly PM_{10} concentrations during the study period, which gives a better understanding of the dispersion of PM_{10} and informs some of the assumptions used in the model, including the construction of homogeneous 'exposure districts' surrounding each of the monitoring sites. That analysis is followed by the construction of the model, including the designation of local sources of PM_{10} and associated emission parameters. Three case studies are presented to demonstrate the use of the resulting model, each considering sub-groups of the population; (1) senior males living in apartments in the center of the city (Bloomsbury) who smoke and cook with gas, (2) females who live outside the center of the city (Brent), work in Bloomsbury, smoke and cook with gas. For this second group, exposures are estimated in the spring (2a) and the summer (2b). The results from these three case studies are analyzed and discussed in the remainder of Section 5, with overall conclusions in Section 6.

2 A Framework for Exposure Estimation

In this section, the general framework for exposure assessment is described. The full description consists of three parts; (i) the background probabilistic structure which provides the 'backbone' of the modelling framework, (ii) an explanation of the 'building blocks' (structural elements) of the model and (iii) the links between these blocks (the stochastic structure). To aid the clarity of the explanation, much of the technical detail is included in Appendices A and B. This section concentrates on the parts required to implement the model, including those which users will need to perform their own analysis using the WWW facility.

2.1 Description of the Model.

This sub-section, together with the associated appendices (A and B) describes the general framework presented by Bauer et al. (1996). The stochastic models used are based on the theoretical probability space, $(\Omega, \mathcal{A}, P)$. The individual components of this space are as follows:

- Ω - the sample space of all underlying information relating to an individual's exposure to the pollutant in question. Items sampled from this space are denoted by ω, a purely conceptual device labelling the sum of information about recognized and unrecognized factors associated with an individual's exposure, for example their residential address
- $\mathcal{A}$ - the collection of subsets of Ω representing the outcome of a sampling experiment. Membership of this class ($A \in \mathcal{A}$) occurs when ω lies in A. Intuitively, $P(A)$ is the fraction of all ωs in A if all elements of Ω have an equal probability of being sampled.

P - the population distribution of the ωs, which is unknown.

In reality, not all the information linked to ω is observed. Instead, only its relevant attributes are measured, leading in general to a random multidimensional array $X = X(\omega)$ of responses. X might be a vector representing exposures in successive time periods. If several pollutant species were being monitored over time, X would be a matrix. Such a matrix could even represent exposure to a single pollutant through several distinct media such as air and water in successive time periods. The proposed setup is therefore general in nature.

The population distribution P, induces a population distribution for X, P_X, defined by $P_X(B) = P(A)$ where $A = \{\omega : X(\omega) \in B\}$. Like P it is unknown and in practice needs to be estimated by sampling. Good sampling designs are vital to obtaining good estimates of population distributions. To characterize such designs imagine $\tilde{\omega}$ is selected at random from Ω. Sampling is 'representative' if the probability that $\tilde{\omega}$ is in A equals $P(A)$ for all A. In that case, the sampling plan is 'unbiased'. That notion extends in an obvious way to the sampling distribution of say $Y = X(\tilde{\omega})$ as well as to the case where a sequence of ωs will need to be drawn to construct sensible estimates of the population estimate. These ideas also extend to sampling from sub-populations when the population is stratified as in Section 4. In this paper sub-population sampling designs are assumed to be unbiased.

Modelling has been simplified (as in Section 4) by splitting ω into two parts, one associated with individual factors (**I**) like 'age' and one with factors external (**E**) to the individual like ambient pollution levels that pervade the study area. Thus, we assume that $\Omega = \Omega_I \times \Omega_E$, or in other words, that $\omega = (\omega_I, \omega_E)$ for each $\omega \in \Omega$ where $\omega_I \in \Omega_I$ while $\omega_E \in \Omega_E$. Further details of the estimation of probabilities within this framework can be seen in Appendix A.

This model has a broad domain applicability in that 'sampling' can be either from a real or a simulated population. The latter may well obtain when ω_E is generated by a complex computer model of environmental processes. Alternatively, the latter could be a small finite population of patterns of individual behavior combined with computer models to yield outcomes of randomized activity patterns. That is exactly the sort of population encountered in pCNEM simulation where ω_E is fixed. Continuing with this example, note that the model includes the possibility realized in pCNEM of choosing different individuals at different time-points. In other words, ω_I can index a composite individual made up of successively sampled individuals, this allows an estimate of the uncertainty associated with each of the estimates to be calculated, in the form of a standard error. Further details of how such a standard error can be obtained can be seen in Appendix B.

2.2 Structural Elements of the Model.

This sub-section contains a description of the building blocks of the model, including the associated processes, such as the interaction between individuals and their environment, that link ω (the theoretical information obtained by sampling) with X (the observed information). Some of these elements may depend on space or time, but for expository simplicity

that dependence is suppressed.

As described in Section 1, the sample space Ω, is split into two parts, comprising internal (Ω_I) and external (Ω_E) factors. This stratification is achieved using discrete stratification variables, S to get $\Omega_s \stackrel{defn}{=} \{\omega : S(\omega) = s\}$ and $\Omega = \cup\{\Omega_s : s \in \mathcal{S}\}$, $\mathcal{S}$ denoting S's range. This stratification: (i) gives us a natural way of incorporating important qualitative variables into the model; (ii) creates homogeneous sub-populations with respect to exposure, thereby facilitating model development.

Although the aforementioned stratification variables will be context dependent, they will fall into either one of the two categories. The collection of internal variables, **I**, consists of individual stratification variables such as age, and are not constant across all the individuals in the study population.

The external variables, **E**, will consist of external stratification variables such as the daily maximum temperature. These strata are constant across individuals. Modelling can now be done conditional on both the **I** and **E** variables.

Human behavior is an important part of any realistic exposure model and a set of variables, **B**, relating to this needs to be defined. The ranges of such variables involve:

MEs: microenvironments - homogeneous personal exposure settings such as kitchens or cars. A random individual's random activity pattern takes him or her through a sequence of MEs during the exposure period. An example used in the BEADS model for benzene exposure is the ME associated with fuelling a car.

As: activities - activities carried out in the MEs. This aspect of behavior can affect the impact of the exposure. Playing tennis or jogging can increase breathing rate and consequently, the dosage of an hazardous air pollutant.

POSs: positions - the geographical locations of the MEs. Activity diaries usually do not record POS. However, it can be of great importance in the presence of a heterogeneous pollution field. In such cases, if it is not modelled then severe measurement error may be introduced into the estimates of exposure. An example from the model proposed by Ott et al (1988), uses a variable denoting whether an individual is 'outdoors and within 100 feet of a road' as a proxy for their position (POS).

A number of additional model components need to be defined. The first, **AF**, denotes the temporally varying random ambient pollution field that takes values at points of a discrete, not necessarily regular, grid distributed over the study area. To date exposure models have only considered concentrations of a single pollutant, however it could be extended to represent multivariate responses, thus enabling several pollutants to be considered simultaneously.

Since AF cannot be completely measured, ambient monitors measure AF at selected points in space and time, leading to, **AM**, the ambient monitoring field. AM can be considered a subfield of AF (if measurement error is ignored) and whilst it is not intrinsic to X, it must be included in the model since it provides the only information available about AF.

Individuals move randomly through the AF, exposing themselves to varying concentrations of a pollutant. However, actual exposure depends on how much AF penetrates the MEs through which an individual moves. For example, ambient ozone can enter an indoor ME only through the exchange of outdoor and indoor air. Furthermore, it depends on the rate at which this reactive gas deteriorates within the indoor environment once it is admitted.

Thus, actual exposure from ambient sources, X_A differs from the AF levels at the position P of the ME. In fact, $X = X_A + X_L$ where X_L denotes exposure from local sources. In the case of carbon monoxide for example, X_L could be a substantial fraction of X as a result of emissions from a gas cooker in a kitchen.

2.3 Stochastic Structure of the Model.

Exposure models such as pCNEM and SHEDS use structural linkages between model components (the structural elements described in the previous sub-section) that are expressed through mathematical equations. In general, the form of these equations are determined using statistical techniques applied to training data, for example, in pCNEM the AM are related to the ME concentrations using such equations to determine X_A. Where applicable the form of the equations can be based on physical laws of the process; in pNEM such a model is used to determine the ozone's decay rate after it has reached an ME's interior.

Whatever their origin, structural linkage models leave residual uncertainty about X. Some of that uncertainty will be due to prediction error and some to replacing model parameters with estimates which will introduce sampling error. This stochastic uncertainty means that the structural parts of the model must have an associated stochastic element, which is at the heart of probabilistic exposure models such as pCNEM. Measuring the uncertainty in X is of fundamental importance, for amongst other things, it will determine the confidence that decision-makers, such as planners, regulators and those formulating air quality objectives can place in the results.

The hierarchical framework for the required stochastic links is described in terms of conditional probabilities. For that description, (T, V), (T| V) and (T) denote generally the joint distribution of T and V, the conditional distribution of T given V and the marginal distribution of T, however expressed, for any random objects T and V.

Assume

$$\begin{aligned}(X, X_L, X_A, B, AM, AF, I, E) &= (X \mid X_L, X_A)(X_L \mid B) \\ &\quad \times (X_A \mid ME, P, AF)(AF \mid E) \\ &\quad \times (AM \mid AF) \\ &\quad \times (B \mid I, E)(I, E), \end{aligned} \tag{1}$$

where B=(A,ME,POS). This model is sufficiently general as to embrace not only pCNEM but earlier computational models such as SHEDS. The uncertainty about $X = X(\omega)$ can now be represented and it can be shown that the population distribution of X is a function of the conditional probability $(X|AM)$. The proof of this, together with further details of

the derivation of the conditional probabilities which are required can be found in Appendix C. One such result shows that for fixed AM the conditional probability, $(X \mid AM)$, is a mixture of (X|I, E, AM), any one of which could hypothetically be used in setting air quality criteria. Alternatively, the criteria could be more targeted and based on (X|I, E, AM) for specified choice of I and E. Progress has been made on the development of the $(AF \mid AM)$ module in a variety of contexts (see, for example, Brown, Le and Zidek 1994; Le, Sun and Zidek 1997; Sun, Le, Zidek and Burnett 1995; Diggle et al 1998). A module for $(ME, A \mid I, E)$ is discussed in Section 3.

The conditional distribution of position given the (internal and external) variables denoted by $(POS \mid I, E)$ is more problematical since, as noted earlier, time-activity diaries will not usually indicate the location of MEs (although postal codes yield it for residential MEs). One solution, relies on a coarse grid $\mathcal{P}$, an accurate specification of a surrogate POS* for POS, but a large measurement error in determining X_A. Another, would rely on a accurate model $AF(POS)$ given the unknown POS and a another model $(POS \mid POS^*)$, given POS*, of position. Which of these strategies is preferable in terms of reducing X_A's measurement error is unclear.

Note that

$$(B \mid I, E) = \int (POS \mid ME, A, I, E, POS^*)(POS^* \mid ME, A, I, E) d(POS^*) \times (ME, A \mid I, E).$$

$(POS \mid ME, A, I, E, POS^*)$ might be determined from such things as census data, traffic survey data, workforce surveys, and maps of school locations within school districts. Without such data, $(POS \mid ME, A, I, E, POS^*)$ might be taken to be uniform over POS^*, the region in which POS lies.

3 Implementing the General Framework.

Sections 1 and 2 provide a general framework for predicting exposure. This section concentrates on certain elements required to implement an exposure prediction model. Specifically, Subsection 3.1 describes the module, $(ME, A) \mid (I, E)$ based on a possible stratification of I and E in the case of PM_{10}. Then in Subsection 3.2, several alternatives for modelling $POS \mid (I, E)$ are discussed from a practical point of view.

3.1 Modelling Human Activities.

In general, the $(ME, A) \mid (I, E)$ module provides a database of activity patterns for use in specific applications of the general framework. For any given pollutant, that data-base will have its diary records organized in a particular fashion. For clarity of explanation, PM_{10} is the target pollutant with the knowledge that other pollutants could be handled in a similar way.

To realistically capture the variations in human behavior needed to predict exposure, the time-activity records of real subjects can be used. The databases derived from the National Human Activity Pattern Survey (NHAPS) conducted in the United States and a similar survey conducted in Canada provide convenient sources of such records. The databases are very similar so they have been integrated into a single database for the purpose of the online computer model described below in Section 4.

The subjects in the NHAPS telephone surveys were asked to recall what happened during the 24 hours of the day before the interview day. Specifically, they were asked 66 major questions:

1 to 7 concerned the household, for example, whether any gasoline or kerosene was stored anywhere in the home;

8 to 39 asked about events during the 24 hour period, including their start and finish times as well as if anyone smoked during that event;

40 to 66 concerned background information such as the year of birth and employment status.

Each diary consists of a sequence of starting times, durations and NHAPS locations from midnight to midnight. There were 78 possible location codes, grouped in 5 categories. In addition to the diary records, the NHAPS surveys include a large number of demographic variables (almost 500), the most important of which appear in Table 1.

Variable	Description
RESPID	Respondent ID
REGION	Region
EMPLOYMENT	Employment status
AGE	Age
SEX	Gender
WORKPLACE	Workplace
YEAR	Year of diary
MONTH	Month of diary
DAY	Day of diary
MAXTEMP	Maximum temperature

Table 1: The most important NHAPS demographic variables.

As noted in Section 1, modelling behavior for subpopulations of the population of interest may be preferable to treating the population as a whole. Such a strategy leads to the stratification of the population by **I**, representing individual strata and E, representing external strata (I $\times$ E strata for short). For PM_{10} , elements of **I** are, 'AGE', 'WORKING-STATUS' and 'GENDER' forming homogeneous sub-populations called DEMOGRAPHIC

GROUPS (DGRPs). These would be selected with reference to available data sources so that in particular, the subpopulation fractions in these groups could be found. These fractions (*i.e.*marginal probabilities) of membership in these DGRPs would then enable population level aggregates to be found from the conditional probabilities given membership in the DGRPs.

Likewise E strata are formed using the stratification (*i.e.* conditioning) variables each measured according to the calendar date of the sampled person-day:

- SEASON: summer or winter
- TEMPERATURE: cool or warm
- DAY TYPE: weekday or weekend.

The temperature classification would be based on the daily maximum temperature on that date, 'cool' being defined as below a 'cut point' that differs in winter and summer. An analysis of the relationship between people's outdoor activities and temperature provides suitable cut points for the temperature classification.

Each (exposure) event reported in an activity diary has the following components:

- POSITION: geographical location;
- ME: microenvironment;
- SMOKE: passive smoking status.

NHAPS defines 'microenvironments' differently than pCNEM (Details can be found in Brauer et al 1996). We discuss the problem of specifying POSITION abbreviated as POS in our notation the next section.

3.2 Modelling $POS \mid (I, E)$ – a Practical Perspective

The problem of specifying the geographical location (POS) is now considered. An ideal activity survey would give POS for each activity so that the associated exposure could be predicted from the $AF(POS) \mid AM$ module (see Subsection 2.3). However, such a survey would not be feasible. Instead, in the NHAPS surveys, 'home' and 'work post code' are the best available indicators of POS, forcing some additional modelling. For example, one might well take $(POS \mid (I, E))$ to be uniform over the smallest census survey 'tract' that can be linked to the postal code. Ambient monitor (AM) measurements would then be interpolated down to the tract level. Then, after stratification and the incorporation of census data the exposure of the whole population of the study area could be estimated.

As noted in Subsection 2.3, this approach has the advantage that AF(POS) and hence X_A can be predicted from POS, while incorporating model uncertainty in specifying $POS \mid POS^*$. (Recall that POS^* was defined above in Section 3.1 as a surrogate of POS.) However,

such an approach ties the eventual results of the exposure study closely to the underlying activity survey designed for a particular area and year. To ensure the validity of population level extrapolation of the results of that study, the activity survey must be spatially and temporally representative of the demographic population subgroups. This requirement is related to the I × E stratification described above. For example, with the PM_{10} stratification enough person-days with different characteristics as indicated by DGRP, SEASON, DAYTYPE, and TEMPERATURE are necessary. To ensure geographical coverage, that number should be multiplied by the number of tracts within the study area. Finally, the association between exposure and the activity-survey implies that an exposure study for a new study year or area will (unrealistically) entail a new activity survey.

An alternate approach to modelling $POS \mid (I, E)$ like that of pCNEM in Section 4, assumes activities can serve as prototypes for use in other situations say with a different area and year. For example, an activity pattern corresponding to an index DGRP = 1 (0-5 kids), SEASON = 1 (winter), DAYTYPE=1 (weekend) and TEMPERATURE =1 (warm) represents a typical activity for 0-5 year old children on a warm winter weekend. Consequently, the position variable defined in pCNEM needs to be as vague as possible. In the current version of pCNEM, POSITION becomes a binary variable, called the 'district identifier'. It tells us only if the activity happened within a home- or work-district. pCNEM links activity pattern, ambient pollutant and the 'cohort' defined by DEMOGRAPHIC GROUP, HOME DISTRICT and WORK DISTRICT. This somewhat coarse approach has the advantage of allowing data from several surveys to be used and allows census data at the aggregated level to be utilised, thus supplying exposure indices for the population of any urban areas for different years.

4 Using the pCNEM Model.

This section describes how the pCNEM model works in practice and gives details of a unique feature, namely the ability for users to develop and run models remotely using the WWW site. Details of how prospective users may resister to use the site and the requirements regarding model parameters, variables and data are given in Appendix D.

4.1 Overview.

The pCNEM simulator basically generates a sequence of pollutant concentrations to which an randomly selected individual is exposed over time. This sequence is termed the *personal exposure sequence*. The generation is a fairly complex stochastic process that follows the randomly selected individual in his activities over the period of the simulation. The individual is thought of as visiting one *microenvironment (ME)* after another as he or she is involved in his or her activities through time. The universe in which the individual evolves is partitioned into a set of *exposure districts* or *district* for short, for each of which the pCNEM simulator will require the ambient pollutant concentration throughout the period of the simulation. Although this partition is arbitrarily fine, in its current implementation,

the pCNEM simulator will position the individual in one of two specific districts, the *home district* and the *work district*, which are selected at the time the simulation is launched.

The generation of the *personal exposure sequence* involves the following arrays:

- $ambient_{d,t}$ the ambient concentration in district d at time t,
- $source_{d,m,t}$ the pollutant production in microenvironment m of district d at time t,
- $local_{d,m,t}$ the concentration to which an individual in microenvironment m of district d at time t would be exposed,
- $d(t)$ the district in which the individual finds himself at time t, and
- $m(t)$ the microenvironment in which the individual finds himself at time t.

The simulator has two major tasks

- the creation of the $local_{d,m,t}$ array, and
- the creation of an activity sequence for the randomly selected individual (the activity sequence will determine $d(t)$ and $m(t)$).

Once the two tasks are completed, it is simply a matter of tracking the individual through his activities in the $local_{d,m,t}$ array. Formally, the result of the simulation is the sequence of personal exposures EXP

$$EXP_t = local_{d(t),m(t),t} \tag{2}$$

Figure 1 shows a $2 \times 3 \times 24$ *local* array and highlights the path followed by an randomly selected individual through time. His activities take the individual from home, to outdoor, to indoor (in the work district), to outdoor, to indoor and back to home at times 5, 6, 14, 16 and 17. The sequence of personal exposures EXP is obtained by reading the *local* array at the highlighted cells from left to right.

The ambient concentrations $ambient_{d,t}$ need to be supplied. The simulator will automatically generate random values for the pollutant productions $source_{d,m,t}$. The many parameters that affect this random generation can each be adjusted. The *local* concentrations are derived from the *ambient* concentrations and the *source* productions

$$local = \Phi(ambient, source) \tag{3}$$

where Φ needs to be determined.

The activity sequence of the randomly selected individual that is being tracked through the *local* array is obtained by concatenating randomly selected NHAPS diary records. It is possible to restrict the random selection of the diary records to those that belong to a particular *demographic group*. A *demographic group* is defined by imposing a restriction of

some of the demographic variables of the NHAPS records The simulator can also be configured to restrict the random selection of diary records to those that match the simulation day in terms of season, day type and meteorological conditions. This ability takes into account the possibility that season, day type and meteorological conditions may affect the activity patterns of individuals.

4.2 The Derivation of the Local Array.

There are two types of microenvironment: *closed* and *open*. A closed microenvironment is one for which the derivation, (see Equation (3)), of the *local* array involves a mass balance equation. Such a microenvironment may have sources that produce amounts of the pollutant and its volume and other quantities are used in the mass balance equation that is used to derive the resulting *local* concentration. On the other hand, an open microenvironment is one for which there is no source and for which the *local* concentration is a simple linear transformation of the *ambient* one.

Closed microenvironments.

A closed microenvironment may have sources of the pollutant under consideration, thereby increasing the concentration to which its occupants are exposed. On the other hand, as air from within is exchanged with air from outside, there is a tendency for the microenvironment *local* concentration to adjust itself to the *ambient* one. The mass balance equation governs the path to this equilibrium

$$\frac{d}{dt}C_{in}(t) = \frac{S(t)}{V} + \nu F_p C_{out}(t) - (\nu + F_d)C_{in}(t) \tag{4}$$

where

- $C_{in}(t)$ is the concentration (*local*) inside the enclosure at time t (mass/volume),
- $C_{out}(t)$ is the concentration (*ambient*) outside the enclosure at time t (mass/volume),
- $S(t)$ is the pollutant generation (*source*) rate inside the enclosure at time t (mass/time),
- V is the volume of the enclosure (volume),
- ν is the air exchange rate (1/time),
- F_p is the penetration factor (unitless), and
- F_d is the deposition (decay) rate (1/time).

The penetration factor F_p and the deposition rate F_d affect the position of the equilibrium while the air exchange rate ν affect the speed at which it is reached. The transformation

Φ in Equation(3) amounts to solving the differential equation (4) and assuming that the inputs *ambient* and *source* do not change too quickly, we have an approximate solution in

$$C_{in}(t+1) = 0 \tag{5}$$

Open microenvironments.

For an open microenvironment, Equation (3) requires the specification of a slope b_m and an intercept a_m. Equation (3) for the *local* array is a simple linear transformation, in other words, Φ is linear in it's first argument:

$$local_{d,m,t} = a_m + b_m * ambient_{d,t} \tag{6}$$

4.3 The WWW Interface.

The following provides a brief guide to a unique feature of the pCNEM system, namely the WWW interface. The interface can be used to configure a model and to launch simulations. Models can be created from scratch using the interface, but it may be easier to use an already existing model as a starting point. The site contains an existing model for PM_{10} exposures in the Vancouver area which is accessible to all users as a template.

The computer model consists of several components:

- the model data,
- the location,
- the simulator,
- the demographic groups,
- the microenvironments, and
- the sources.

The model data component is used to supply the ambient field and possibly the meteorological data corresponding to the area and the period for which the simulation will be done. The location component is used to define the mapping from NHAPS locations to the microenvironments defined in the simulation. The simulator component is used to launch simulations. The demographic groups component is used to edit and inspect the demographic groups defined for the model. The microenvironments component is used to edit and inspect the microenvironments defined for the model. The sources component is used to edit and inspect the sources defined for the model.

Prospective users of the pCNEM system should apply to *jim@stat.ubc.ca*. Once registered, users can

- upload their own databases;
- add their own 'micro-environments (MEs)', key building blocks for the model;
- add sources interior to each ME;
- fit appropriate source parameters
- make replicate runs of the model remotely
- download outputs in spreadsheet format to their own PC for further analysis and application.

More detail is available on-line by accessing the 'help' screens.

5 pCNEM Predictions of PM_{10} for London, 1997.

This section demonstrates how the model described in Section 4 has been used to find conditional predictive exposure distributions for PM_{10} in London in 1997. This is done by considered three specific case studies each referring to a particular subpopulation of Greater London (Sub-section 5.2). These sub-populations were selected for their potential interest and diversity. Moreover, under realistic assumptions, their membership totals can be estimated.

Thus in principle, population level distributions could be found by combining these conditional exposure probabilities with those for the other subpopulations, $\{GL_i\}$ of Greater London using the standard formula, $P(B) = \sum P(B|GL_i)P(GL_i)$. For example, B could be the event that in 1997, the 98th percentile of the daily average PM_{10} exposures for a randomly selected Greater Londoner exceeds 50 $\mu g\ m^{-3}$. The subpopulation GL_1 could be that considered in the first case study of Subsection 5.2, namely males older than 60 years, who smoke and live in an apartment equipped with a gas stove.

The aggregation of subpopulation probabilities will be left for future work. Instead, this paper focuses on finding subpopulation probabilities to illustrate pCNEM's use. These analyses demonstrate pCNEM's flexibility and in particular consider the effect of reductions in the ambient concentrations of PM_{10} which might be affected by changes in policy. The effects of such reductions, known as 'rollbacks' can be assessed in terms of the changes in exposure experienced by individual members of the population. In addition, the effect of using different emission parameters can be set for different sub-populations (GL_is) is assessed, for example, the difference in the distributions of the number of cigarettes smoked by randomly selected males in GL_1 and in the 0-to-17 year age range, respectively, can be accommodated.

Site Name	Site Type	Easting	Northing
Bexley	Suburban	5518	1763
Bloomsbury	Urban Centre	5302	1820
Brent	Urban Background	5200	1840
Eltham	Suburban	5318	1816
Haringey	Urban Centre	5266	1791
Hillingdon	Suburban	5348	1862
North Kensington	Urban Background	5321	1831
Sutton Roadside	Roadside	5256	1640

Table 2: Description of the PM_{10} monitoring sites used in the model.

5.1 London's PM_{10} Field in 1997

Daily data from 8 PM_{10} monitoring sites were uploaded to the pCNEM site together with maximum daily temperatures for London for 1997. Details of the pollution monitoring sites are given in Table 2. Each site represents a pCNEM's exposure district, that is a geographical area surrounding it whose ambient PM_{10} level can realistically be imputed to be that of its central monitoring site. Previous work on the spatial distribution of PM_{10} in London for this period (Shaddick and Wakefield, 2002) indicates that the PM_{10} field is relatively homogeneous, meaning that the boundaries of the exposure regions are less critical than might otherwise be the case. In order to compute the marginal membership probabilities as described in the introduction to this section, the regions would be linked to census wards for which the required population counts are routinely available. Summary statistics for the 8 monitors in this study are given in Table 3.

	Total	% Missing	Mean	SD	25%	50%	75%	Max
Bexley	8760	7.4	23.0	15.8	13	19	28	243
Bloomsbury	8760	3.9	26.6	16.2	17	23	32	349
Brent	8760	4.9	21.6	14.4	13	18	26	235
Eltham	8760	9.5	21.2	13.8	13	18	25	224
Haringey	8760	3.2	26.4	16.2	16	22	31	292
Hillingdon	8760	2.6	25.0	16.7	14	21	32	227
N.Kensington	8760	2.5	24.3	15.7	15	20	29	219
Sutton	8760	0.9	24.1	14.5	15	21	29	219

Table 3: Summary of hourly PM_{10} measurements at eight sites in London during 1997. Units are μgm^{-3} (except for % missing). The minimum value in all cases is 1 except for the last two sites where it is 2.

Only a relatively small proportion of measurements are missing even though a small number of negative numbers and zeros (for which there were no meaningful interpretation) were excluded. The mean exceeds the median (the 50 percentile) at all sites, pointing to the

heavy right distribution tail seen typically in air pollutant concentration series, suggesting the log-normal distribution might be appropriate for modelling purposes. For this reason, the (natural) log scale was used in some of the plots below.

The similarity of the values in each column of Table 3, indicates the great similarity in the series for the eight sites. The high correlations observed between the individual series from the sites, was exploited when imputing missing hourly values, where the average value for that hour at the remaining sites was used.

Figure 2 shows the high level of spatial homogeneity. In the upper panel (a), the hourly measurements from the most central of all the eight sites (Bloomsbury) are plotted against time and can be compared to the average of the corresponding measurements from the remaining seven sites. Little difference is observed between the two series, as was the case in the corresponding figures using the other sites as the 'central' site (not shown). The high correlation between the Bloomsbury and the seven site average series is seen in the scatterplot, 2(b). Again, similarly high correlations were seen when looking at the other sites in turn (not shown).

The eight time series of hourly values exhibit relatively little systematic variation. For example, the weekly effect (geometric mean of all the hourly measurements over all eight sites for that week) plotted on a log scale in Figure 3(a) shows no clear seasonal pattern and only small week-to-week fluctuations. To examine the daily effect, shown in Figure 3(b), the geometric average of all the hourly measurements over all sites and all weeks $1, \ldots, 52$ were calculated, after removing the weekly effect. For day 1 in week 1, all $24 \times 8 = 192$ hourly measurements for day 1 (Sunday) from all sites, were divided by week 1's average. Measurements for each of the days $2, \ldots, 7$ were treated analogously and in turn measurements for week $2, \ldots, 52$. The resulting, adjusted, average concentrations were plotted on the log scale. Note that '1' on the vertical scale indicates no deviation from the weekly effect while 1.05 means a 5% deviation above that effect. A distinct day-of -the-week is seen, consistent with the smaller volumes of traffic seen on Sunday in London. However, the absolute variation is small. Finally, in Figure 3(c) hourly effects (hour 1 being the period from 0:00 to 01:00 hours) are shown, after adjusting for the week and day effect in a manner analogous to that used for computing the day effect. Once again a distinct pattern of (small) variation is seen.

After completing the initial analysis, the PM_{10} data file, with imputed missing values was uploaded to the pCNEM site.

5.2 Case Studies

In this example, five microenvironments (MEs) are incorporated into the model: Outdoors, Home, Transit, Indoors not at home, and Bar/Restaurant. These are the MEs incorporated by Özkaynak et al (1995) in their study. Home and Bar are closed and the remainder, open. Thus, parameters for the mass balance equations are needed for the former, regression

coefficients for the latter.

The ME sources of PM_{10} emissions are: 'home.cooking', 'home.smoking', 'home.other', 'bar.smoking', and 'bar.other'. pCNEM links sources to MEs so they can have different parameters (that may also depend on population subgroup). The proposal of Özkaynak et al (1995) and Burke et al (2001) is adopted, that of including 'other' as source of unaccounted for emissions. Note that any of these sources can be excluded simply by setting their emission levels to 0 rather than deleting them from the model. Home is the primary ME since so much of peoples lives are spent therein.

The parameters in the mass balance equation in Section 4 are based on published data. These data yield the elements of Table 4. The distributions for the air exchange rate are based on the study reported by Murray and Burmaster (1995). In particular, results are used for what they call 'Region 1', that part of the US that seems to best approximate London. Obviously, data for London would improve the model's accuracy. The distributions for residential volume are drawn from Murray, D (1997) based on a survey of US housing. Again, if available, data for London would be preferable. The remaining entries in Table 4 are based on Özkaynak et al (1996) and the well know PTEAM study. Note that, as in Özkaynak et al (2001), parameters can be specified as probability distributions to account for the uncertainty about their exact value.

While Table 4 gives emission distributions for cigarettes, estimates of the average number of cigarettes smoked in the closed MEs are also required. The number of cigarettes is assumed to follow a Poisson distribution, for which an estimate of the mean is required. Members of a subpopulation of smokers will smoke some of these cigarettes themselves while they are in those MEs. For both smokers and non-smokers, environmental tobacco smoke will potentially be produced by other smokers in those MEs. These need to be incorporated into the Poisson mean.

Table 5 gives the weekly number of cigarettes consumed by smokers in the UK in 1998, the year closest to 1997 for which data are available. Ökaynak et al (2001, Table 7) provide estimates of the number of cigarettes smoked by others in a residential ME. From there we are able to compute the means presented in Table 6. Although the age ranges in these two surveys do not match, they can be combined to yield the approximate hourly means reported in Table 7 thus giving the required parameter for the Poisson distribution in home.smoking source. To obtain this mean it is assumed that the cigarettes are consumed between 07:00 and 22:00 hours.

Ökaynak et al (1995) and Ökaynak et al (2001) treat the Bar/Restaurant as an open ME. Linear regression predictors are used to predict the ME concentrations in both cases. In the first of these two references, different predictors are used to predict ME concentrations from ambient concentrations differently for smoking and non-smoking establishments. In the second, a single regression predictor is used but a third term is incorporated involving a random covariate representing the active smoking count. This assumes a uniformly distributed random simultaneous number of between 0 and 3 cigarettes at any one time. For the purpose of this demonstration, 'Bar', loosely refers to bars and restaurants, and is defined as a closed environment. A small non-random sample of pubs in London carried out

Table 4: Distributions Assumed for Residential Mass Balance Equation. 'Winter' means {Dec, Jan, Feb} and each of the remaining seasons follow in 3 month blocks. LN(μ,σ) denotes the lognormal distribution with mean μ and standard deviation σ *in logarithmic space*. N(μ,σ) denotes the normal distribution with mean μ and standard deviation σ.

Parameter	Categories	Distribution
Air exchange rate (h^{-1})	Winter	LN(-0.8, 0.7)
	Spring	LN(-1.2, 0.8)
	Summer	LN(0.2, 0.7)
	Autumn	LN(-1.2, 0.5)
Volume (m^3)	Detached	LN(6.02, 0.50)
	Attached	LN(5.78, 0.48)
	Apartment	LN(5.24, 0.43)
	Other	LN(5.34, 0.36)
Penetration (Unitless)		N(1.00,0.08)
Deposition (h^{-1})		N(0.68,0.21)
Cigarette emission (μ g /cig)		N(22, 4.23)
Cooking emissions (μ g /min)		N(4.1, 0.79)
Other emissions (μ g /hr)		N(5.6, 1.56)

Table 5: Average daily cigarette consumption by smokers aged 16 and over, by age and gender, England 1998. Source: Office of National Statistics General Household Survey, 1978 to 1998.

	All	16-19	20-24	25-34	35-49	50-59	60 & over
Men	15	10	14	13	17	17	15
Women	13	10	11	12	15	15	12

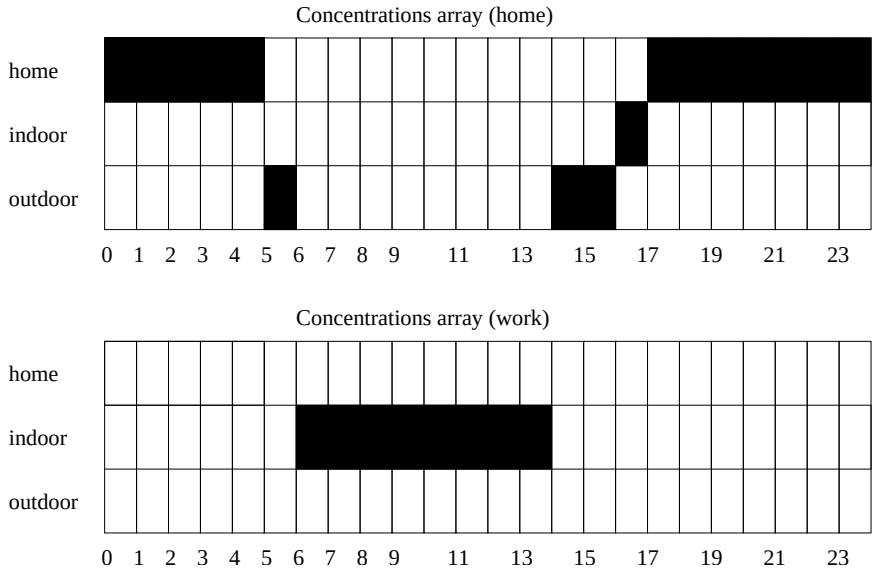

Figure 1: The path followed by an randomly selected individual through 24 hours of time. The 2 $\times 3 \times$ 24 local array categorises their exposure to pollution at home or work and whether they are exposed to sources either in the home, (other) indoor or outdoor.

Table 6: Mean number of cigarettes smoked by other inhabitants per day in a residence by smoking status, age and gender.

Smoking Status	Age Group	Male Mean	Female Mean
Smoker	12-17 yr	15	2
	18-64 yr	14	3
	$\geq$ 65 yr	17	4
Non-smoker	12- 17 yr	14	3
	18-64 yr	11	2
	$\geq$ 65 yr	12	2

Table 7: Mean number of cigarettes smoked per hour in a residence by smoking status, age and gender.

Smoking Status	Age Group	Male Mean	Female Mean
Smoker	12-17 yr	1.67	0.80
	18-64 yr	2.00	1.07
	$\geq$ 65 yr	2.13	1.07
Non-smoker	12- 17 yr	0.93	0.20
	18-64 yr	0.73	0.13
	$\geq$ 65 yr	0.80	0.13

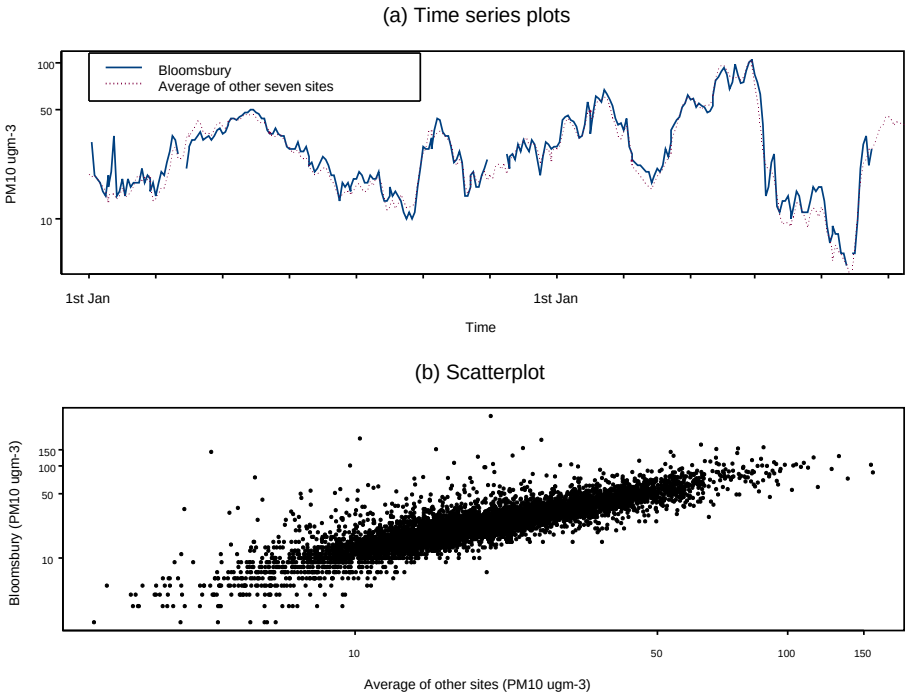

Figure 2: (a) Time series plots for hourly Bloomsbury PM_{10} and that of the average of the remaining seven sites. (b) The corresponding scatterplot for the two series.

by the first author and his assistant, found between 11 and 32 patrons in a pub, or about 22 on average. The number of smokers at any one time ranged from 0 to 6 (about 1.9 on average). That number was between 0 % and 27% of the patrons or 8.5% on average.

Looking at this issue in a different way, the Office of National Statistics General Household Survey 1998, indicates that about 28% of England's adult population are smokers. According to Table 5 (and assuming 15 hours in the 'day'), each of these will smoke about 1 cigarette per hour. Thus, $22 \times 0.28 = 6$ smokers would be expected among the average of 22 patrons. They would together smoke about 6 cigarettes in an hour. A cigarette takes about 6 minutes to smoke. So at any given time, cigarettes ignited at most 6 minutes before or after that time will be burning. Assuming (tenuously) that smokers behave independently of one another, a simple probability calculation suggests that $12/60 \times 100 = 20\%$ of these cigarettes would be burning at any one time during that hour, that is about 1.2 cigarettes. This number is in reasonable agreement with the 1.9 calculated from the non-scientific survey as well as the 0 to 3 assumed in the SHEDS model. Thus in the bar.smoking source model, we assume the random number of cigarettes has (for convenience) a Poisson distribution with a mean number of 6 cigarettes smoked per hour during the periods, 12:00-15:00

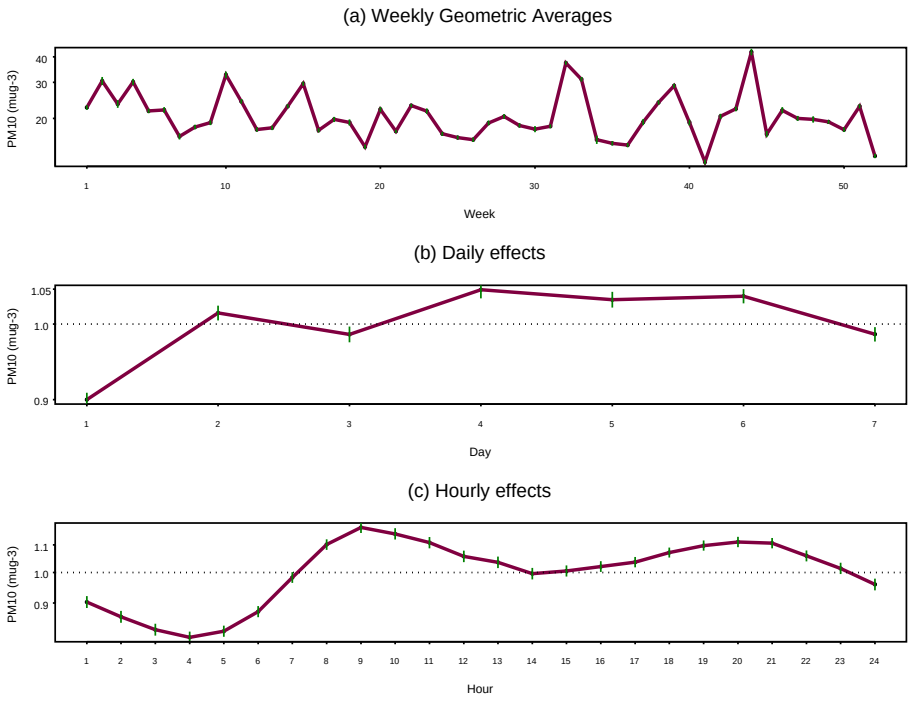

Figure 3: Systematic variation in the joint 1997 hourly time series of all eight London PM_{10} sites. The 52 week effects, 7 day effects and 24 hour effects are shown in panels (a), (b) and (c) respectively. Note that in (b) and (c), due to the use of logged values, 1 on the vertical scale means no deviation from the baseline.

and 18:00-23:00 hours. The other parameters needed for that ME are adapted values of the corresponding residential parameters.

The intercept β_0 and slope β_1 for the model that links ambient levels with concentrations in the open MEs now needs to be determined. Obviously, for the Outdoor ME, $\beta_0 = 0.0$ and slope $\beta_1 = 1.0$. The remainder are found using a method suggested by Ökaynak et al (1995) and the coefficients for $PM_{2.5}$ in Ökaynak et al (2001, Table 3).

The idea is that $PM_{2.5}$ is an important and sometimes primary fraction of PM_{10} , depending on the emission source. Thus, $PM_{2.5} = \alpha\ PM_{10}$ for some fraction, α. Ökaynak et al (1995) take $\alpha = 0.6$ (albeit for public and commercial buildings). Then given a model ME:$PM_{2.5} = \beta_0 + \beta_1$ ambient:$PM_{2.5}$, we obtain α ME:$PM_{10} = \beta_0 + \beta_1\ \alpha$ ambient:PM_{10}, or in other words, ME:$PM_{10} = \alpha^{-1}\beta_0 + \beta_1$ ambient:PM_{10}. Note that the slope does not change whatever the value of α. In contrast, the intercept is inflated 1.67 times if $\alpha = 0.6$ in going from the $PM_{2.5}$ model to one for PM_{10} .

Values from this analysis and the values given in Ökaynak et al (2001) are given in Table 8. For the Indoor Non-residential the values given for Office, School and Store in that cited reference have been averaged.

Table 8: Coefficients for the ambient to the open ME link models.

Microenvironment	β_0 ($\mu g\ m^{-3}$)	β_1 (unitless)
Outdoor	0.0	1.0
Indoor Non-residential	10.8	0.5
Car	16.3	1.0

With no London time-activity data available for implementing pCNEM, the coefficients used by Brauer et al (1996) are used with days classified as 'warm' if the daily maximum temperature is above $13°C$ in the winter and above $29°C$ in the summer.

In the case studies that follow, two population subgroups are considered, with the second case considering exposure distributions separately in both the spring and summer of 1997. In each case, 30 replicate pCNEM runs were used to estimate those distributions. One very interesting feature of pCNEM is the ability to reduce the ambient PM_{10} levels to any specified level (by 20% in these examples) under a hypothetical abatement program. The predictive distributions are re-estimated under that scenario.

Case Study 1: Senior Male. This case study considers the sub-population of males who are over 64 years of age, smoke, live in an apartment with a gas cooker in the Bloomsbury exposure district. The study period is defined as 'Spring', the months of March-May 1997. The individual is assumed to be non-working so that by default his work and study districts are identical.

Case Study 2: Working Female in Spring and Summer. This sub-population consists of working women who smoke, live in Brent in semi-detached dwellings that use gas as the cooking fuel and work in the Bloomsbury exposure district. Two sub-cases are

considered, one covering the spring and the other the summer of 1997.

For each of the cases above, the output from a single pCNEM run consists of a sequence of 'exposure events' for a randomly selected member of the subpopulation under consideration. That individual is composite; a different time-activity is selected for each succeeding day. This composite individual better represents his/her subpopulation's activity patterns than any single member would do. The theoretical basis for this approach is set out above in Subsection 2.1. Each event in the sequence takes place in a random micro-environment and consists of exposure to a randomly varying concentration of PM_{10} for a random number of minutes.

The output from the pCNEM model, consisting of 30 replicates generated for each of the three cases, can be analysed in a variety of ways. The starting point here is to look at the average instantaneous exposure in each of the five microenvironments used in model. Thus, for each run, day, and microenvironment, the product of event time and event concentration is divided by the total amount of time spent in that microenvironment. By calculating the average over days for the daily values computed in that way, the average 'instantaneous' PM_{10} level for that microenvironment can be estimated. The 30 replicates give the estimated predictive distributions, as is shown by boxplots for each of the three cases in Figure 4

Comparing the upper and middle panels of that figure for seniors and working women, both reflecting random exposure experiences in the spring of 1997, it can be seen that:

- for microenvironments other than 'bars', the greater boxplot widths for females than seniors may reflect greater variety in their time activity studies;
- their in-home exposures are very similar, even though the greater volumes of the homes of females should have diluted indoor source emissions more;
- the similarity of their non-home, 'indoor' exposures is expected since the females work in Bloomsbury where the seniors live. Hence, if, as might be expected, the seniors tend to remain in their home district during the day, their indoor exposures would be similar;
- the greater variability of exposure in the 'outdoor' microenvironment would also be expected since the females are commuting from one district (Brent) to another (Bloomsbury). That should naturally increase variability in their outdoor concentrations over those of seniors.
- the same factor may help explain the comparatively higher levels of seniors in transit over those of working women, if the former remain in their exposure district.

The middle and lower panels of Figures 4 enable a comparison of exposures of the single subpopulation of working females addressed by the second case, during the spring and summer of 1997. Since lifestyles generally defer between these two season, one would expect to see differences in exposure that go beyond any differences in ambient pollution levels.

In fact, little difference is observed from one season to the next, suggesting no substantial differences in time-activity patterns of women in this subpopulation again indicating the lack of strong seasonality in London's PM_{10} field in 1997 as seen in Figure 3.

While the average instantaneous concentration of PM_{10} in a microenvironment gives the level of PM_{10} typically encountered there, sojourn times seem more important in terms of potential impacts of that level. Figure 5, provides comparisons resembling the last ones that take weight concentrations by time of exposure. More precisely, the weighted daily product of concentration by duration in minutes of exposure event is calculated and divided by 60 minutes to get the average over days of hourly cumulative exposures. This in effect moves from averages over days of one minute cumulative exposures (*i.e.* approximate 'derivative' of the cumulative concentration function known as 'instantaneous' concentration'), to the corresponding averages for hours. In terms of this cumulative index of exposure, 'home' tends to be higher for the senior males than the working females in spring. However, for working females in summer, that index rises well above the one for senior males. That observation suggests that these women spend proportionally more time in the home during the summer months than they do in the spring.

In terms of cumulative exposure, 'bar' contributes little in all cases, although for women, we see more variability in this measure than that for senior males.

Since the females in the case study are working they might be expected to experience higher levels of cumulative exposure than that for senior males and the figures confirm this. They also agree with our a priori expectations that this measure would show higher levels in spring than summer, the latter being lower when women are on summer vacations, for example. In fact, based on our previous observation for women in the 'home' microenvironment, it may be inferred that women tend to move from the office to home when they go on vacation.

Working females in summer seem to have similar cumulative exposures outdoors as senior males do in spring, that measure being higher in both cases than that for working females in spring. However this exposure is small compared with that for all other microenvironments except for bars. That suggests either that ambient levels are low when people are outdoors or that they spend little time there. In contrast, the large values of that measure of cumulative exposure for 'transit' compared to one for 'outdoors', suggests travel takes place when ambient levels are high (since travel times tend to be relatively small). However, senior males do not seem to sustain such high levels of that exposure as do working women.

In order to examine how exposure patterns over the day, the sum over exposure events was computed for each hour of each day of the products of the concentrations times duration, divided by 60 minutes. The result was then average for each hour over days. For each replicate run, this gives an average hourly concentration. The 30 replicates in turn, allow the predictive distribution for that effect to be estimated. For the cases under study, the results appear in Figure 6.

It can be seen that these hourly exposures starting rise dramatically around 07:00 hours along with their variability (that tends to be small during the night). For the senior males,

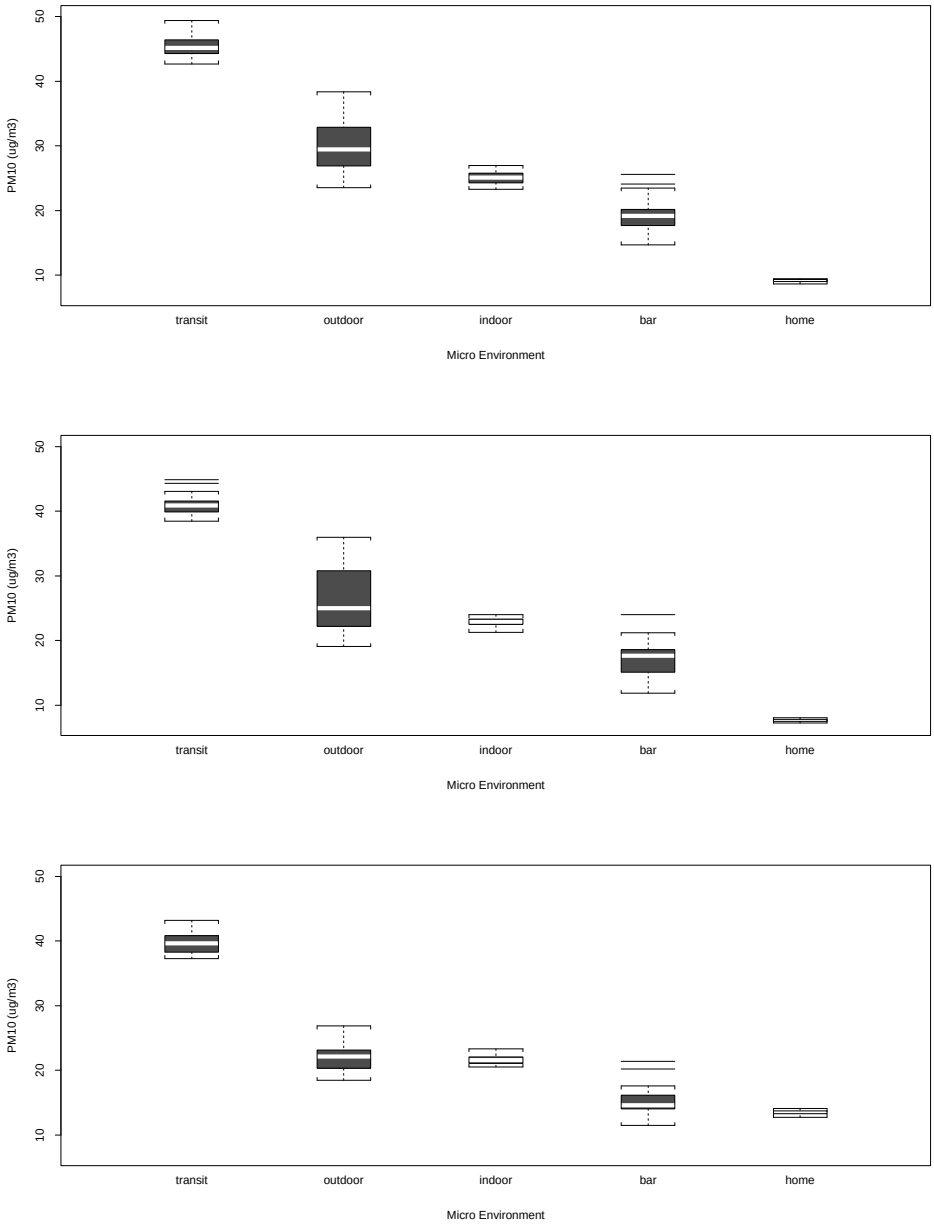

Figure 4: Boxplots depict the estimated predictive distributions of PM_{10} exposure in 1997 for a random member of a selected subpopulation of Londoners. Here distributions are for instantaneous average exposures by microenvironment. The upper panel is for springtime exposures of senior males living in apartments in Bloomsbury, who smoke and cook with gas. The other two panels are for Women who live in semi-detached Brent dwellings, Work in Bloomsbury, smoke, and cook with gas. The middle and bottom refer respectively to spring and summer for those woman.

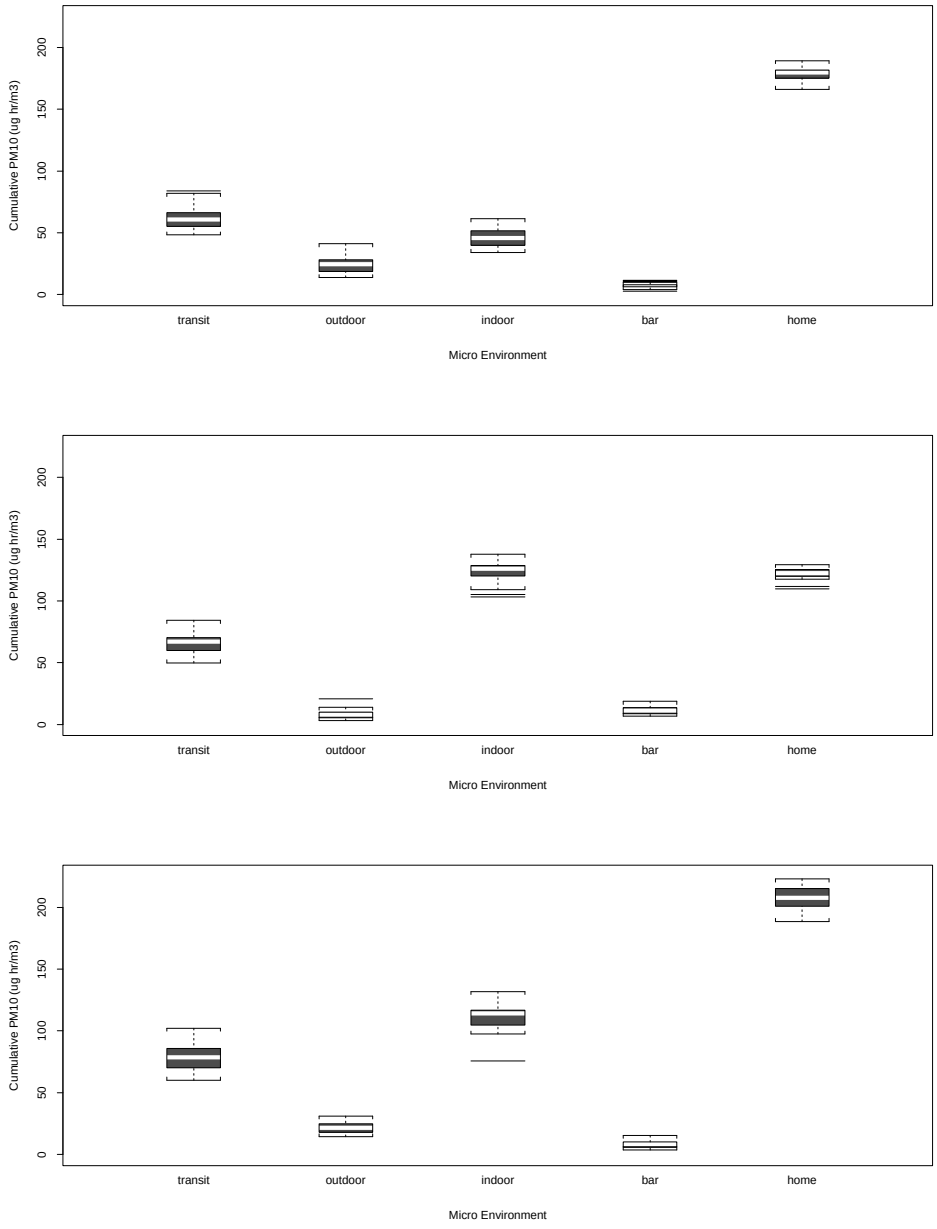

Figure 5: Boxplots depict the estimated predictive distributions of PM_{10} exposure in 1997 for a random member of a selected subpopulation of Londoners. Here distributions are for cumulative one hour average exposures by microenvironment. The upper panel is for springtime exposures of senior males living in apartments in Bloomsbury, who smoke and cook with gas. The other two panels are for Women who live in semi-detached Brent dwellings, Work in Bloomsbury, smoke, and cook with gas. The middle and bottom refer respectively to spring and summer.

a peak of around 20 μgm^{-3} is reached around 10:00 hours, while for working women these peaks are reached around 16:00 hours with values around 22 or 23 μgm^{-3}. Exposures for the senior males drop steadily from their peak to about 10 μgm^{-3} around midnight. In contrast, those for working females have a broad peak spanning several hours. In spring, those exposures also drop to around 10 μgm^{-3} while in summer they bottom out at close to 15 μgm^{-3}. That points to a higher ambient level at night in Brent where these women are supposed to live. We next make comparisons over days for the sub-populations under study. Specifically, for each replicate, we compute the daily averages of the hourly values computed above. The results for daily averages appear in Figure 7.

Both the senior males and working females experience similar high exposures for a period of about 4 days. However, that for women exhibits greater variability during that period and, in particular, includes a number of replicates where high levels of average daily exposure are seen, approaching or exceeding 50 μgm^{-3}. This is still well below the standard for daily average level for PM_{10} of 150 μgm^{-3} in both the UK and the US.

During August of 1997, similar peaks are seen in daily average exposures of working females. Again, there are notable extremes among the replicate values, at about or exceeding 50 μgm^{-3}.

Finally, Figure 8 shows the differential impact of a hypothetical 20% deflation termed 'rollback' of actual PM_{10} levels in the spring of 1997. The capacity of pCNEM to enable such 'scenario' analyses proves to be one of the programs most important features, giving regulators a way to check the impact of proposed changes on subpopulation groups. The deflated scenario values are set by the user and computed as the linear reduction, $baseline + p \times (x - baseline)$ for every hourly datum, x, for both Brent, the residential exposure district for the working female, and Bloomsbury where that female worked and the senior male lived. In our example, the reduction factor and baseline were chosen somewhat arbitrarily to be p=0.8 and baseline = 15 μgm^{-3}.

The senior males are regarded as more susceptible to morbidity from high levels of PM_{10} . Moreover, the Figure shows that they will enjoy greater benefit from the hypothetical reduction than the working females. This differential benefit may well be due to the male's greater duration of outdoor activity than the working female's during periods when the pollution levels are highest.

However, the size of the differential is small compared with others seen, for example, by using the Vancouver template supplied with pCNEM for a child living in Burnaby versus a male who lives in Burnaby and works in Vancouver. The child, a member of another susceptible group, enjoys a substantial differential benefit compared to the working male. This is because levels of PM_{10} in Burnaby tend to be high compared to those in Vancouver where the working male will spend a sizeable fraction of his workdays.

6 Discussion and Conclusions.

This paper has presented an online exposure estimation platform and illustrated how it can be readily adapted for use in estimating exposures for areas different to that on which it

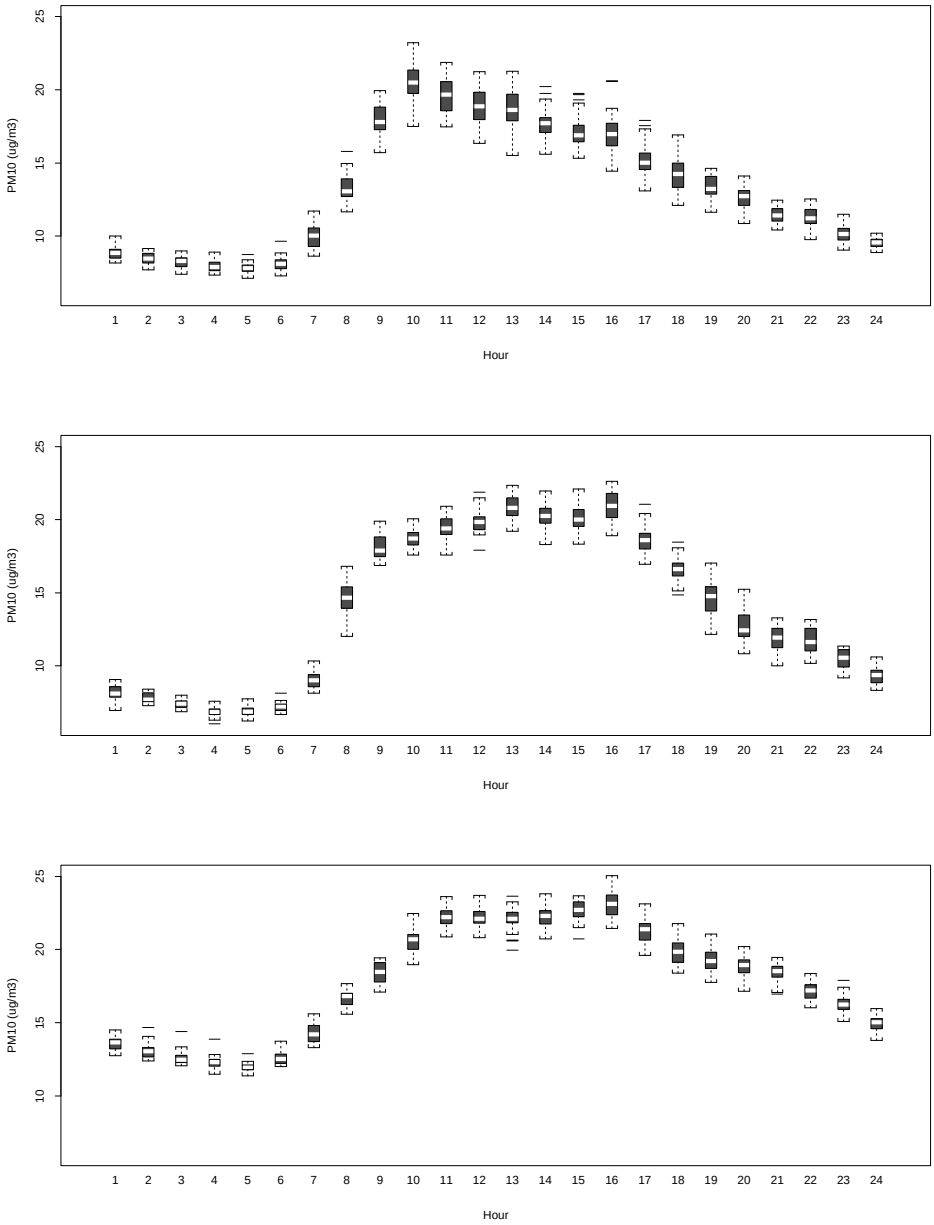

Figure 6: Boxplots depict the estimated predictive distributions of PM_{10} exposure in 1997 for a random member of a selected subpopulation of Londoners. Here distributions are for hourly exposures. The upper panel is for springtime exposures of senior males living in apartments in Bloomsbury, who smoke and cook with gas. The other two panels are for Women who live in semi-detached Brent dwellings, work in Bloomsbury, smoke, and cook with gas. The middle and bottom refer respectively to spring and summer.

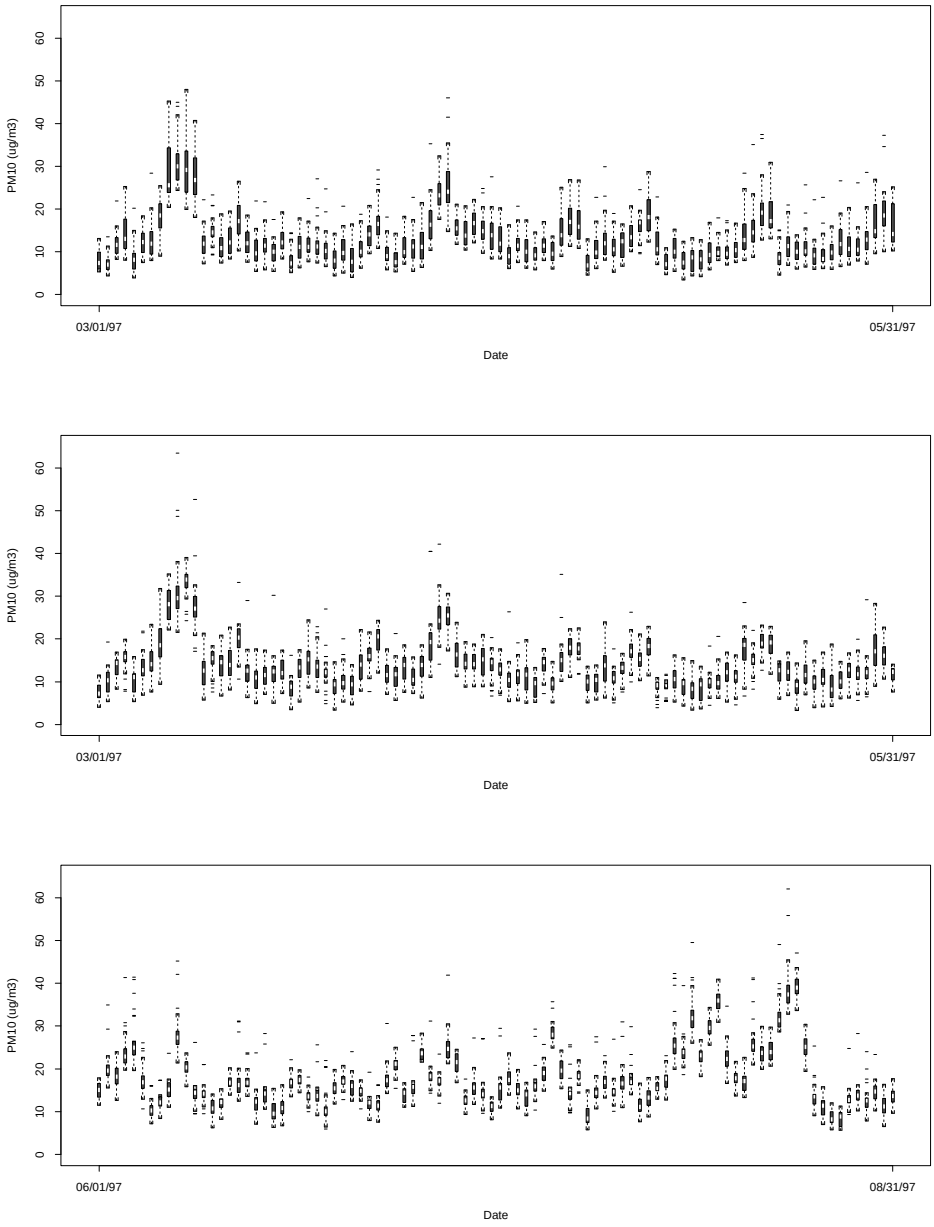

Figure 7: Boxplots depict the estimated predictive distributions of PM_{10} exposure in 1997 for a random member of a selected subpopulation of Londoners. Here distributions are for average daily exposures. The upper panel is for springtime exposures of senior males living in apartments in Bloomsbury, who smoke and cook with gas. The other two panels are for Women who live in semi-detached Brent dwellings, Work in Bloomsbury, smoke, and cook with gas. The middle and bottom refer respectively to spring and summer.

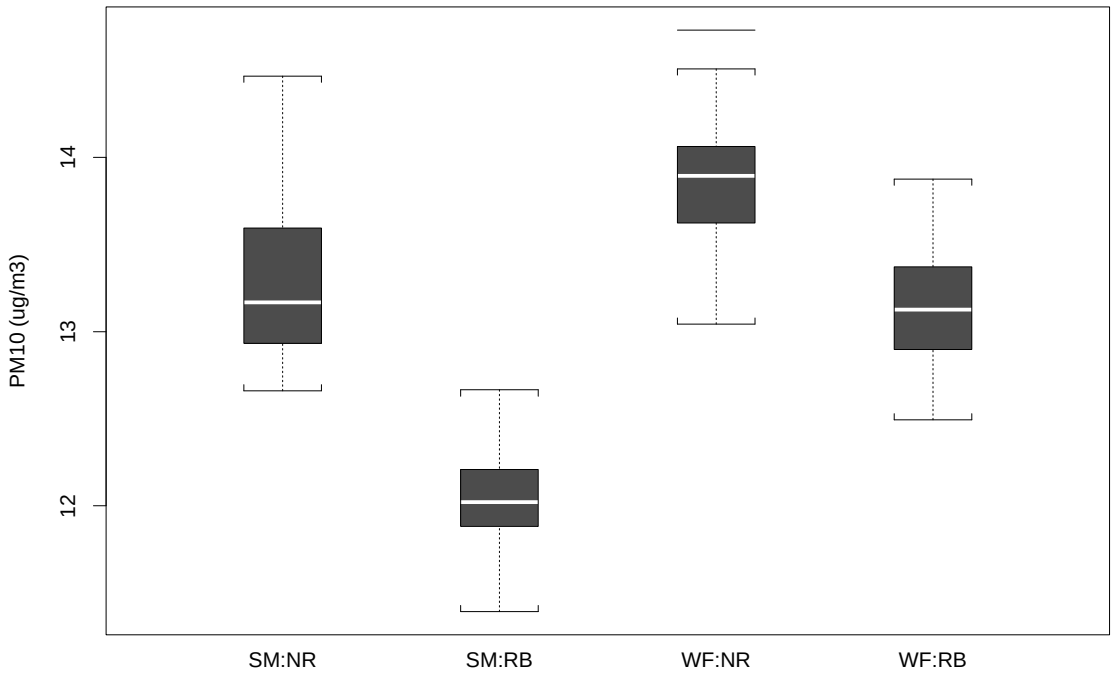

Figure 8: Boxplots compare the estimated predictive distributions of PM_{10} exposure in 1997 for a random member of a selected subpopulation of Londoners. Here distributions are for average daily exposures on spring days for senior males (labelled 'SM') and working females (labelled 'WF') before ('NR') and after ('RB') a 20% rollback in hourly levels. Both subpopulations are smokers and use gas as a cooking fuel. The males live in a Bloomsbury apartment, the women in a semi-detached Brent dwellings. However the latter work in Bloomsbury.

was originally designed. In this paper, exposures were successfully estimated for randomly selected Londoners in various designated sub-populations. The model executes quickly, is flexible, and may be run remotely from the user's PC.

To maximize the platform's flexibility, it is not programmed to compute marginal exposure predictive distributions by averaging over London's sub-populations. However, this would be of interest to policy-makers, after all, some sub-populations whose members might be heavily impacted by high PM_{10} exposures may at the same time be negligible in size, making that impact minimal. The platform could readily be extended to compute such weighted averages, although this would add additional complexity with small benefit. This would require a complete set of population counts for the sub-populations to be uploaded by the client to enable the host server in computing those weighted averages. As such, the calculation would be more efficiently carried out by users locally after downloading the output from their pCNEM model simulations.

In the example presented, data for London was used wherever available, further enhancements could be made by incorporating a time-activity data file for Londoners along with better parameter values for the various micro-environment emission sources.

The model needs to be tested empirically as a predictor. That needs to be left to future work as resources become available. At the same time, an application to environmental epidemiology has been started, the goal of which is to produce more realistic estimates of the exposure to pollutants, particularly within susceptible groups, which will then be used to assess the effects on health. Part of this work will be to compare the estimates from the pCNEM model with those directly available from the monitoring sites, which in the past have used in the majority of epidemiological studies into the short-term effects of air pollution.

Acknowledgements.

We are indebted to Dr Sandra McBride for comments that helped improve the quality of our presentation. Part of the work in this paper was completed with the help of a Fellowship from the Engineering and Physical Sciences Research Council of the United Kingdom, while the first author was a visitor in the Department of Mathematical Sciences at the University of Bath. The work was completed while he was a visitor at the Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina. He is indebted to both of these institutions for generously providing facilities.

References.

Burke, JM, Zufall, MJ, and Ökaynak, H (2001). A population exposure model for particulate matter: case study results for $PM_{2.5}$ in Philadelphia, PA. *J Exposure Analysis and Enviromental Epidemiology*, 11, 470-489.

Brauer, M, Chatfield, C, Le, ND, Zhang, H and Zidek, JV. (1996). Generalized space-time models of pollution exposures whose members interact with their environment.

Unpublished report.

Brown, PJ, Le, ND and Zidek, JV (1994). Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics*, 22: 489-509.

Diggle, PJ, Tawn, JA and Moyeed, RA (1998). Model-based geostatistics (with discussion). *Appl Statist*, 47, 299-350.

Hartwell, TD et al (1984). Study of carbon monoxide exposure of residents of Washington, DC and Denver, Colorado. EPA-600/54-84-031, US Environmental Protection Agency, Research Triangle Park.

Law, PL, Lioy, PJ, Zelenka, MP, Huber, AH, and McCurdy, T (1997). Evaluation of a Probabilistic Exposure Model Applied to Carbon Monoxide (pNEM/CO) Using Denver Personal Exposure Monitoring Data. *J Air Waste Management Assoc*, 47, 491-500.

Le, ND, Sun, W, and Zidek, JV (1997). Bayesian multivariate spatial interpolation with data missing-by-design, *Journal of the Royal Statistical Society, Series B*, 59, 501-510.

MacIntosh, DL, Xue, J, Ozkaynak, H, Spengler, JD, and Ryan, PB (1995). A population-based exposure model for benzene. *J Exposure Anal Environ Epidemiol*, 5, 375-403.

Murray, D and Burmaster, D (1995). Residential air exchange rates in the United States: empirical and estimated parametric distributions. *Risk Analysis*, 15, 459-465.

Murray, D. (1997). Residential house and zone volumes in the United States: empirical and estimated parametric distributions. *Risk Analysis*, 17, 439-445.

Ott, W, Thomes, J, Mage, D and Wallace, L (1988). Validation of the simulation of human activity and pollutant exposure (SHAPE) model using paired days from Denver, CO carbon monoxide field study. *Atmos. Environ.*, 22, 2101-2113.

Özkaynak, H, MacIntosh, D, Sue, J and Zhou, H (1995). Predicted distribution of personal exposures to PM_{10} in Canada. Contract Report. Environmental Health Center. Health Canada.

Özkaynak, H, Xue, J, Spengler, J, Wallace, L, Pellizzari,I and Jenkins, P (1996). Personal exposure to airborne particles and metals: results from the particle team study in Riverside, California. *Jour Exposure Analysis and Environmental Epidemiology*, 6, 57-78.

Robinson, JP and Thomas, J (1991). Time spent in activities, locations and microenvironments: a California-national comparison. Environmental Monitoring Systems Laboratory, Las Vegas, Nevada.

Shaddick, G and Wakefield, J (2002). Modelling multiple pollutants and multiple sites. *Appl Statist*, 51, 351-372.

Sun, W, Le, ND, Zidek, JV and Burnett, R (1995). Bayesian multivariate spatial interpolation: application and assessment. Technical Report No. 146, Department of Statistics, University of British Columbia.

Appendix A

Estimating population distributions

As described in Section 2, the sample space of all possible information relating to an individuals exposure to pollution, Ω is split into two parts, representing individual (I) and external (E) factors, i.e. $\Omega = \Omega_I \times \Omega_E$ and $\omega = (\omega_I, \omega_E)$ for each $\omega \in \Omega$ where $\omega_I \in \Omega_I$ and $\omega_E \in \Omega_E$

Assume that $\mathcal{A}$, the collection of subsets of Ω representing the outcome of a sampling experiment, includes all events of the form $\mathrm{A}_I \times \mathrm{A}_E$ where A_I and A_E are events in Ω_I and Ω_E, respectively.

Given $\omega_I = \tilde{\omega}_I$, for any given $\tilde{\omega}_I$, the (conditional) population distribution of X, the observed information, is $P(X \in B \mid \omega_I = \tilde{\omega}_I)$ for all B. A standard result from probability theory tells us that

$$P_X(B) = E_I[P(X \in B \mid \omega_I = \tilde{\omega}_I)]$$

where E_I denotes the average over all $\tilde{\omega}_I$ in Ω_I with respect to the population distribution. This shows that $P(X \in B \mid \omega_I = \tilde{\omega}_I)$ would, if known, be an unbiased estimator of $P_X(B)$. When several successive $\tilde{\omega}_I$s are sampled in an unbiased manner, we readily see that any weighted average of the conditional probabilities, if known, yield an unbiased estimator.

Similar reasoning shows that the conditional probability distribution given $\omega_E = \tilde{\omega}_E$, if known, also yields an unbiased estimator of $P_X(B)$. However, the last conditional probability for a fixed $\tilde{\omega}_E$, itself proves of interest in pCNEM's scenario analysis to determine the impact on population level exposures as a result of an intervention. Since in practice it is not known, it is estimated (in pCNEM) by repeated (unbiased) sampling of ω_Is.

As a generalization of the case of estimating $P_X(B)$ considered above, consider any real-valued function of X, say $G(X)$. Note that the special case obtains when we let $G(X) = I\{X \in B\}$, the 0-1 indicator function of B. The population averages $[E(G(X)]$ and $[E(G(X) \mid \omega_E = \tilde{\omega}_E]$ can then be expressed in terms of population and conditional population distributions, respectively. Their associated unbiased estimators are then readily seen.

Appendix B

Estimating the standard error associated with exposure estimates

By obtaining a measurement of the uncertainty associated with the exposure estimates, the model can be used in statistical inference. Assume each of N individuals in a finite population are exposed to a random pollutant concentration at n successive times to yield

$X = (X(t_1), ..., X(t_n))$. To calculate the expected number, ν, of person-times among these N individuals and n times whose individual exposures exceed a specified level, x^o, let X^{ik} be the exposure of individual k on day i for all i and k. Then $\nu = E[\sum_{i=1}^{n}\sum_{k=1}^{N} I\{X_{ik} > x^o\}] = \sum_{i=1}^{n}\sum_{k=1}^{N} p_{ik}$, where $p_{ik} = E[I\{X_{ik} > x^o\}]$. Note that that if $p_{ik} = p$ for all i and k, we obtain an expression involving the familiar formula for the expectation of the binomial distribution, namely $\nu = nNp$.

To estimate ν using the pCNEM sampling scheme, suppose a random individual K_i is sampled on day i where $\pi_k = P(K_i = k),\ k = 1, \ldots, N$ depends on the sampling design used. Then $\hat{\nu} = \sum_{i=1}^{n} I\{X_{iK_i} > x^o\}/\pi_{K_i}$ is an unbiased estimator of ν. When, in particular, $\pi_k = N^{-1}$ for all $k = 1, \ldots, N$ we obtain $\hat{\nu} = Nn^+$ where n^+ denotes the number of days among the n for which that selected individuals had an exposure exceeding x^o. The pCNEM methodology uses precisely this estimator albeit in the case where ω_E is held fixed.

Our model explicates and justifies the use of that estimator and thus enables a standard error $se_{\hat{\nu}}$ to be calculated. To represent that standard error consider the following additional notation:

$$\begin{aligned} \bar{p}_i &= N^{-1}\sum_{k=1}^{N} p_{ik} \\ \bar{p}_{j|i} &= N^{-2}\frac{\sum_{k=1}^{N}\sum_{k'=1}^{N} E[I\{X_{ik} > x^o\}I\{X_{jk'} > x^o\}}{\bar{p}_i}. \end{aligned}$$

Intuitively $\bar{p}_i$ represents the average probability of exceedance at time-point i over all individuals, $\bar{p}_{j|i}$ the average conditional probability of exceedance at time-point j given an exceedance at i. With this notation,

$$N^{-2}se^2_{\hat{\nu}} = \sum_{i=1}^{n}\left[\sum_{k=1}^{N}\frac{p_{ik}}{N^2\pi_k} - \bar{p}_i^2\right] + \sum_{(i,j):i\neq j}\bar{p}_i(\bar{p}_{j|i} - \bar{p}_j). \tag{7}$$

In the special case where $\pi_k \equiv N^{-1}$ we find in particular that

$$N^{-2}se^2_{\hat{\nu}} = \sum_{i=1}^{n}\bar{p}_i(1-\bar{p}_i) + \sum_{i\neq j}\bar{p}_i(\bar{p}_{j|i} - \bar{p}_j). \tag{8}$$

Observe that positive correlation between all time-points i and j of exposure exceedances makes $\bar{p}_{j|i} - \bar{p}_j > 0$ and hence the standard error larger as might be expected on intuitive grounds. The standard error will be large as well unless the $\{\bar{p}_i\}$ are not all close to 0 or all close to 1.

Unless exposure exceedances at different time-points are independent we cannot estimate the standard error given above, there being then just one replicate for each $(i,\ j)$ pair of time-points. The need for such replicates is an important consideration in the design of simulated sample methods for exposure estimation.

Appendix C

Determining the required conditional probabilities

The joint distribution of the structural components of the model can be expressed as follows;

$$\begin{aligned}(X, X_L, X_A, B, AM, AF, I, E) &= (X \mid X_L, X_A)(X_L \mid B) \\ &\quad \times (X_A \mid ME, P, AF)(AF \mid E) \\ &\quad \times (AM \mid AF) \\ &\quad \times (B \mid I, E)(I, E), \end{aligned} \tag{9}$$

where B=(A,ME,POS).

Various distributions of interest can now be defined in terms of this joint distribution. The uncertainty about $X = X(\omega)$, that is, its population distribution, can be expressed (assuming ω is sampled in an unbiased way);

$$\begin{aligned}(X) &= \int_{X=X_L+X_A} (X, X_L, X_A, B, AM, AF, I, E) \\ &\quad \times d(X_L, X_A, B, AM, AF, I, E) \qquad (10) \\ &= \int_{X=X_L+X_A} (X \mid X_L, X_A)(X_L \mid B)(X_A \mid ME, P, AF)(AF \mid E) \\ &\quad \times (AM \mid AF) \\ &\quad \times (B \mid I, E)(I, E) d(X_L, X_A, B, AM, AF, I, E) \\ &= \int_{X=X_L+X_A} (X \mid X_L, X_A)(X_L \mid B)(X_A \mid ME, P, AF) d(X_L, X_A) \\ &\quad \times (AF \mid E) dAF (B \mid I, E) d(B)(E, I) d(I, E) \\ &= \int_{X=X_L+X_A} (X \mid B, AF)(AF \mid E) dAF \\ &\quad \times (B \mid I, E) d(B)(E, I) d(I, E) \\ &= \int_{X=X_L+X_A} (X \mid I, E)(E, I) d(I, E), \qquad (11) \end{aligned}$$

where

$$(AF \mid E) = \int (AF, AM \mid E) dAM = \int (AF \mid AM, E) \times (AM \mid E) dAM$$

independently of all other variables above. Note that the uncertainty about the I and E-strata from which ω was drawn has now been incorporated.

pCNEM finds the population distribution in accord with the distribution above, but the distribution actually computed is $(X \mid AM)$ for a fixed sequence of monitoring values. To represent this conditional probability distribution, using Equation (10) developed differently:

$$(X) = \int_{X=X_L+X_A} (X, X_L, X_A, B, AM, AF, I, E)$$

$$
\begin{aligned}
& \times d(X_L, X_A, B, AM, AF, I, E) \\
= & \int_{X=X_L+X_A} (X \mid I, E, AM) \\
& \times (I, E \mid AM) d(I, E)(AM) d(AM) \\
= & \int_{X=X_L+X_A} (X \mid AM)(AM) d(AM). \qquad (12)
\end{aligned}
$$

Thus, (X) is a mixture of (X|AM).