ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

# Analyzing feasibility of setting air pollution levels in India by using Bayes and RVM

[1] Dr.R.Viswanathan, [2]V.VarnaVishakar, [3]T. Edison
[1] Associate Professor, [2][3] Assistant Professor
[1][2][3] GalgotiasUniversity

**Abstract: The safe levels of NAAQS pollution standards have been set but people do not know whether the levels set could be much safe for them with respect to the particular country. This research was to determine the pollution level with respect to NO2, SO2 and PM10using naive Bayes and Relevance Vector Machine algorithms. The steps must be taken not only to reduce pollution but also eradicate it through careful methods of safer waste disposal in all the sectors which aggravates pollution including**

**Keywords – CPCB (Central Pollution Control Board), NAAQS (National Ambient Air Quality Standards), PM10 (Particulate Matter), RVM (Relevance Vector Machines).**

## 1. INTRODUCTION

The healthy atmosphere is determined by the quality of air and water we have in our area. taking only into the account that the country has stepped into industrial revolution without caring that how much the environment gets polluted and how much the health conditions of the people will be affected due to pollution is quite not acceptable by any citizen of the country, people living in a country do not demand to cut industrial revolution but they demand to use safer methods to dispose harmful gases released by industrial sectors and reduction of vehicular pollution. Since petroleum contributes the second largest asset of every country the alternative to the fuel was not put into larger use so far. The Bayes and RVM algorithms were adopted To check pollution levels in all major parts of India so that in coming future strict measures should be taken for reducing pollution by safer disposal of aerial waste. Adequate assessment of environmental exposures that vary within communities in population-based epidemiologic studies is limited by the expense involved in obtaining Measurements at multiple locations, often for prolonged periods. An example is air pollution, for which studies of chronic health effects have traditionally relied on continuous measurements made at central site monitors. Although successful for demonstrating initial associations, central-site measurements fail to capture the large variability in exposures within communities that occurs near roadways and stationary sources . Recent studies have shown that within community exposure gradients may be associated with larger health effects than the between-community exposures used in earlier studies. This transition to studies using within community expo-sure gradients raises measurement and statistical issues. In particular, local-scale monitoring information is needed to calibrate and confirm exposure assignments, and there is increased potential for measurement error in estimated exposure. With the large uncertainty in exposure estimates, questions remain about the validity of results from health effects studies that use exposure surrogates based on incomplete information, such as road buffers or models fitted with sparse monitoring data. The combination of large initial health effects and the heightened potential for errors has prompted researchers to identify the development of models for assessing air pollution exposure within cities as a priority for future research.

## 2. REVIEW OF LITERATURE

The issues regarding global warming, greenhouse effect came into picture as soon as the industrial revolution has started. CFC-11, 12, 13 when released into air and this will warm up the environment.

The CO2 level statistics however not present or not disclosed to public by many governments one must tend to move towards personal calculations towards the measurements of the CO2 and other pollutant concentrations in air. The increase in the atmospheric concentration of carbon dioxide (CO2) and other harmful gases is generally accepted as the main contributor to the anthropogenic greenhouse effect.Likewise SO2, NO2, PM10 which are considered as some of the harmful gases by NAAQS has to be kept in check to maintain a healthy atmosphere.

The capacity to reduce pollution was present in trees and due to deforestation the soil fertility and air quality will get affected in large.Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Classification predicts categorical labels, the co-called "classes", while prediction models are used for forecasting continuous valued variables. Typically, both classification and prediction analyses are a two-step process. In the first step, classification/prediction algorithm is applied on the available data and a decision model is extracted.

The mined decision model encapsulates the knowledge lying in the data in a form such as a decision tree, a neural network, a regression model, a support vector machine, etc. This step is usually called model training phase. The second step is the testing phase, which involves the application of the decision model on data for making decisions (predictions). This phase focuses on testing the ability of the decision model to approximate data not used in training. In this respect, both classification and statistical prediction algorithms have been applied for function approximation in several, yet diverse, domains. To distinguish the terms "classification" and "prediction", the reader may have in mind the following working definition, Where as classification predicts qualitative variables, prediction forecasts quantitative ones.

## 3. Need for setting strict NAAQS limits on industrialized countries

Air is the source of almost all diseases and has a capacity to infect a living organism as it breathes. The environment gets polluted to a considerable extent which is approximately a point less than that of the NAAQS safety level and labelling that particular region is strictly following NAAQS safety standards cannot be fully acceptable.

## 4. MATERIALS AND METHODS

We propose an alternative method to address these two measurement error problems in exposure assessment:

Estimating ‘‘true’’ long-term exposure from flawed short-term measurements and 2) estimating missing measurement (or other covariate) information in locations where no mea-surements have been made. We will utilize Bayesian Markov chain Monte Carlo estimation methods (17) that model the process through which the unobserved exposures are esti-mated. Markov chain Monte Carlo methods can fit an entire multilevel model as a unit, properly taking into account parameter estimation uncertainty at each level of the model.

### *4.1 Data*

We illustrate the methods using data from the Southern California Children’s Health Study. In the Children’s Health Study, continuous, long-term central-site measurements of air pollution were made in multiple communities. Two sea-sonal short-term household-level measurements were made at a subset of study participant residences within communities to characterize local deviations from the community-specific control site measurements at the same times. Using the proposed methods, we estimated household level long-term residential exposures for persons with and without seasonal exposure measurements, and these estimates were used to evaluate the effect of air pollution on lung function in children. This outcome has been well studied by use of central monitors for ecologic, between-community compar-isons (1, 4, 13, 18, 19), but there has been little study of the effect of residential exposures that vary within communities. The above-mentioned estimation procedures were per-formed in a unified Bayesian framework.

### *4.2 Model*

The aims of this paper are to model the determinants of local variation in outdoor concentrations of nitrogen dioxide in the Children’s Health Study in relation to traffic patterns and to use this model to estimate long-term nitrogen dioxide exposure. These estimated determinants will be used as covariates in a model for lung function. In this context, ni-trogen dioxide serves as a proxy for local traffic pollution exposure. Our approach to this problem is grounded in the statistical literature on exposure measurement error.

Measure-ments were made during two seasons, denoted j ¼ 1, 2. The lung function measurements are denoted Yci, the observed household-level exposure measurements Zcij, and the un-observed annual household-level concentrations Xci. Here, we let the subscript * indicate the average over the corre-sponding index. We let Wci denote household-level expo-sure variables that influence local concentrations, such as distance from the nearest freeway and predicted nitrogen dioxide concentration from the CALINE4 (20) line-source dispersion model. The CALINE4 model incorporates dis-tances from traffic densities on all major nearby roads along with the frequency distribution of wind speeds and direction. Personal covariates that influence lung function directly, such as smoking, asthma, and respiratory illness at the time of lung function measurement, are denoted Vci. Our analytical framework consists of the following three submodels, hereafter called the disease, exposure, and measurement models, respectively.

We also model community-level random effects for the dis-ease and exposure models as Ac ¼ x0 þ x1Pc* þ hc and Bc ¼ b0 þ b1Pc* þ kc. Here, we let eci, fci, gcij, hc, and kc represent normally distributed error terms with standard deviations re, rf, rg, rh, and rk, respectively.

Equation 1 models the effect of household-level long-term nitrogen dioxide exposure along with the effects of various personal-level covariates on lung function. Equation 2 uses various household-level covariates to help predict the long-term nitrogen dioxide exposure. Since we only have household-level nitrogen dioxide measurements from two time periods throughout the year for the 233 homes in the study, we model the long-term level of nitrogen dioxide exposure for individual i in community c, Xci, using equation 3. This modeled household-level long-term nitrogen dioxide exposure will simply be referred to as ‘‘modeled’’ nitrogen dioxide exposure in the rest of this paper, but it should be understood to estimate long-term exposure to nitrogen dioxide in the home.

### *4.3 Bayesian estimation procedures*

We set our model in a Bayesian framework and estimate parameters using the Markov chain Monte Carlo method, Gibbs sampling (17). One advantage of these procedures is that missing data can be handled in a natural way. In this technique, each parameter in the model is sampled from its full conditional distribution, that is, the distribution obtained by conditioning on all the other unknowns in the model. Parameters, missing covariates, and latent variables are, in a Bayesian context, seen as random variables, each of which can be estimated using Bayesian parameter estimation techniques.

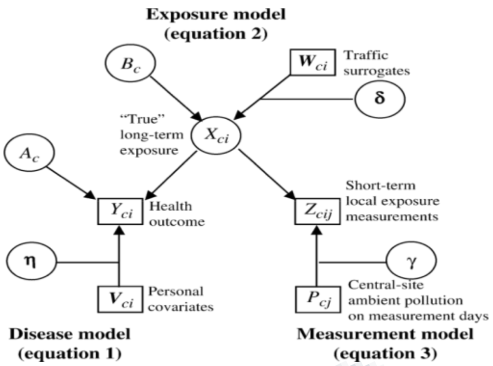


***FIGURE 1. Schematic representation of the overall model used to analyze data in the Children's Health Study pilot project, conducted in the year 2000. Note that, while not represented in the graph, the random intercepts $A_c$ and $B_c$ are modeled by use of central-site exposure measurements $P_c$*.***

The model is updated by use of all available data and current updates of all other parameters in the model. In other words, the relation between various facets of the model such as Y, X, and Z can be obtained for the data that are available, and these relations can be used to impute values for covariates that are not available.

All Markov chain Monte Carlo analyses were conducted using the WinBUGS software package (23). (This program is available upon request from the first author of this paper.) The Bayesian models were run for a burn-in of 20,000 iter-ations, followed by 100,000 iterations that were stored to compute posterior distributions. Diffuse priors were used on all parameters. Regression parameters were assigned N(0, sN) priors (here, sN denotes precision), with sN ¼ 1.0e – 12. Vari-ance components were given flat uniform priors, U(0, sU) as suggested by Gelman et al. (24), as opposed to conjugate priors. We used sU ¼ 100 to define our vague prior for all variance components. Throughout the analyses, all mea-sures of nitrogen dioxide, both estimated and observed, dis-tance to the nearest freeway, and CALINE4-predicted nitrogen dioxide, as well as the outcome, Yci, were measured on a log scale to satisfy the normality assumptions of the models.

Prior sensitivity was examined by making the priors tighter, that is, less vague. For instance, the effects of using larger values of sN (e.g., sN ¼ 1.0e – 6) and smaller values of sU (e.g., sU ¼ 10) were examined. In general, the results were quite robust to changes in these parameter values as long as the prior specifications were sufficiently vague. The sampler exhibited good convergence properties as well.Time-series plots of posterior parameter quantities indicated that the mixing of the sampler was extremely good. Multiple chains were run using different starting values, but the end results were nearly identical for all chains. These features indicate that the total of 120,000 iterations used was much more than necessary to achieve convergence.

The interpretation of parameter estimates obtained when using these kinds of log-log models corresponds to what is commonly known in the regression literature as elasticity. The coefficient in front of a particular covariate is interpreted as the percent change in the response, Y, corresponding to a 1 percent change in the value of the covariate, X, assuming that everything else in the model is held constant. For example, if the model is log(Y) ¼ b0 þ b1 log(X) þ e, then if b1 ¼ 0.2, a 10 percent increase in the value of X will lead to a 2 percent change of 0.02 in the response Y.

The challenge in using the above model setup is to find a way to predict the modeled measures of nitrogen dioxide exposure, Xci in model 5, from observed seasonal measures, Zci in model 4, in such a way that uncertainty in the estima-tion process is properly taken into account. We will compare our previously described Bayesian approach with three frequentist regression approaches. In all three approaches, we model nitrogen dioxide exposure in the second stage model (equation 5) by using the fitted values obtained from the first-stage model (equation 4). The three approaches are de-scribed below.

Comparison of results from these models serves as a useful benchmark for our Bayesian methods. We chose to estimate these frequentist regression models using the SAS software package. This way, we can compare the re-sults obtained from our unified Bayesian model with those obtained from a standard frequentist analysis using a stan-dard statistical software package.

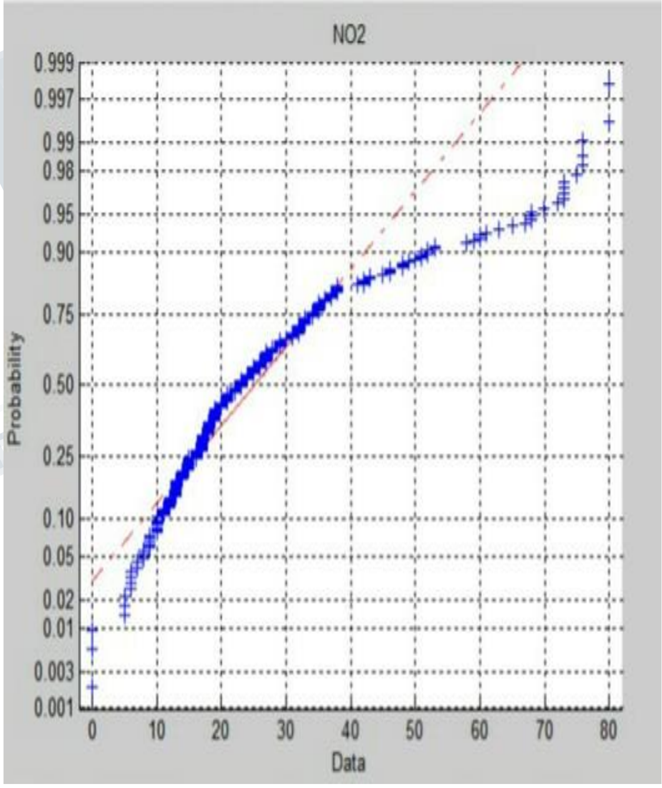
***4.3.1 Using machine learning algorithms Bayes, RVM to classify the dataset***

The study has been extended to entire India and among 246 test sites which monitor pollution statistics in India, the year-wise datasets were gathered from the CPCB which consists of ambient air quality level and concentration levels of NO2, SO2,and PM10in air. Machine learning algorithms were used to classify them to predict which pollutant is present in large quantity and adverse effects caused by these pollutants.The dataset provides the final concentration value for the above three pollutants with respect to the year. The calculations were derived based on various sources like the number of industries present in that particular area, their share among the pollution and vehicular emission levels, dust etc.

$$y = f(x|\mu, \alpha) = \left(\frac{1}{\alpha\sqrt{2\pi}}\right) e^{\underline{\xi}}$$

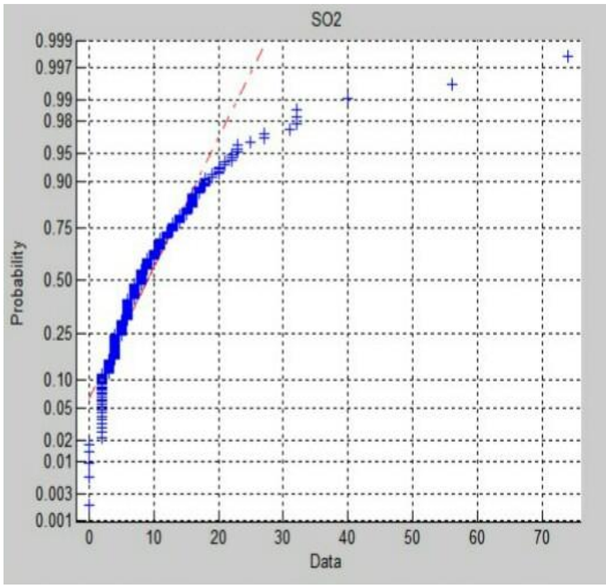
**5. $NO_2$ (Nitrogen Dioxide)**

have to be kept under check and the excess levels of NO2 in air cause several breathing problems and irritation in eye. Bayes algorithm was used to classify the on which range the concentration of the gas is in excess. The results were as follows.



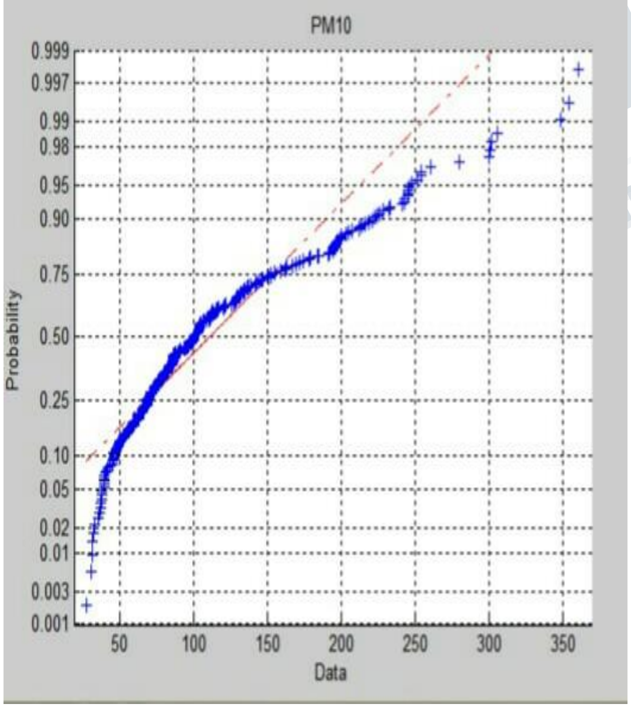
***4.3.2 $SO_2$ (Sulphur Dioxide)***

Burning of coal and oil inside a fuel causes the emission of sulphur dioxide. Excess amount of the gas causes breathing and cardiovascular problems. Acidification of lakes, reduced visibility, corrosion of buildings are also some of the adverse effects of excess concentration.

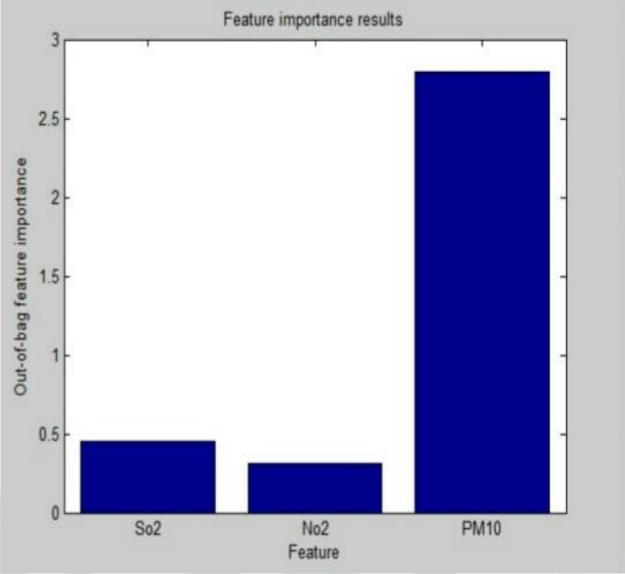


### 6. PM$_{10}$ (Particulate Matter)

The particulate matter mainly a combination of smaller particles has the capacity to reach the lower regions of the respiratory tracks. Damage to lung tissues and other respiratory problems were some of its health impacts.



## 7 . TOTAL CONCENTRATION LEVEL



## 8. PROPOSED ALGORITHM TO DETERMINE POLLUTION LEVEL IN INDIA

The proposed algorithm states the newer way of determining how much quantity of pollution was released in air by various sources. Taking into the account only the total number of sources emitting pollutants and the quantity of pollutants emitted in ppm and also the total quantity of the pollution treated with respect to particular area we can determine the approximate amount of pollutant level concentration in air at the particular point of time. If the government fix the pollution level restrictions based on this value also in addition to the limitations on total quantity of individual pollutant generated one can expect more efficient way of estimating pollution levels several other sources contributing to pollution should also be included in the scenario in accordance to India like open let drains, quantity of untreated garbage in roads, pollution due to tannery etc. the more we analyse the sources of pollution, more we can accurately predict the pollution index.

```
Structstruct_name
{
floatpollutant_source;
}struct_reference;
Func float counter_measures(param1 float
countermeasure_name1, param2 float
countermeasure_name2 )
```

```
{
pollution_treated=countermeasure_name1+countermeasur
e_name2;
returnpollution_treated }
main()
{
struct_reference
total_concentrationof_pollutant_in_air=struct_ref.pollutan
t_source-pollution_treated;
}
```

## 9. RESULTS

The effects of modeled nitrogen dioxide on lung function for FVC are statistically significant and that the effects of nitrogen di-oxide on FEV1 are marginally significant. Figure 2 displays the posterior distributions for the parameter a, the effect of modeled nitrogen dioxide exposure, Xci on lung function for both FVC and FEV1, and demonstrates that modeled nitro-gen dioxide exposure clearly affects lung function in the negative direction. For FVC, the probability Pr(a > 0) equals p ¼ 0.019. Similarly, for FEV1, p ¼ 0.032. (These quantities can be thought of as Bayesian one-sided p values.).The distance to the nearest freeway provides little informa-tion in estimating the level of modeled nitrogen dioxide exposure when CALINE4-predicted nitrogen dioxide is in-cluded in model 2, since distance to freeway is one of the factors included in the calculation of CALINE4-predicted nitrogen dioxide. However, the coefficients take the ex-pected sign, as an increase in distance from a freeway is associated with a decrease in predicted modeled nitrogen dioxide exposure. The results pertaining to the effect of modeled nitrogen dioxide exposure on lung function obtained from the Bayesian model are the only ones that show significance when FVC is used as an outcome. Moreover, Bayesian credible intervals pertaining to this modeled nitrogen dioxide effect, the effect of primary interest, are narrower than all the confidence intervals obtained using frequentist approaches. It is also important to note that results obtained from the frequentist models relating to the effects of traffic-related covariates on seasonal nitrogen dioxide exposures were calculated without regard to the measure of lung function used in the second-stage model 5. Consequently, the results here are the same regardless of the measure of lung function used. The Bayes-ian model uses a unified approach in estimating model.

## REFERENCES

[1] WJ. Gauderman, E. Avol, F. Gilliland , et al. The effect of air pollution on lung development from 10 to 18 years of age. N Engl J Med;351:1057–67, 2004.

[2] DW. Dockery, CA. Pope, X. Xu, et al. An association between air pollution and mortality in six U.S. cities. N Engl J Med;329:1753–9, 1993.

[3] R. McConnell, K. Berhane , F. Gilliland, et al. Air pollution and bronchitic symptoms in Southern California children with asthma. Environ Health Perspect;107:757–60, 1999.

[4] M. Raizenne, LM. Neas, AI. Damokosh, et al. Health effects of acid aerosols on North American children: pulmonary func-tion. Environ Health Perspect;104:506–14, 1996.

[5] M. Jerrett, A. Arain, P. Kanaroglou, et al. A review and evalu-ation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol;15:185–204, 2004.

[6] G. Hoek, B. Brunekreef, S. Goldbohm, et al. Association be-tween mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. Lancet;360:1203–9, 2002.

[7] MM. Finkelstein, M. Jerrett, MR. Sears et al. Traffic air pollution and mortality rate advancement periods. Am J Epidemiol; 160:173–7, 2004.

[8] P. Nafstad, LL. Haheim, T. Wisloff, et al. Urban air pollution and mortality in a cohort of Norwegian men. Environ Health Per-spect;112:610–15, 2004.

[9] B. Brunekreef, ST. Holgate et al. Air pollution and health. Lancet;360:1233–42, 2002.

[10] M. Brauer, G. Hoek, P. Vliet, et al. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. Epidemiology;14:228–39, 2003.

[11] K. Berhane, WJ. Gauderman, DO. Stram, et al. Statistical issues in studies of the long-term effects of air pollution: the Southern California Children’s Health Study. Stat Sci 2004;19:414–19, 2004.

[12] JM. Peters, E. Avol, W. Navidi, et al. A study of twelve southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity. Am J Respir Crit Care Med;159:760–7, 1999.

[13] JM. Peters, E. Avol, WJ. Gauderman, et al. A study of twelve southern California communities with differing levels and types of air pollution. II. Effects on pulmonary function. Am J Respir Crit Care Med;159:768–75, 1999.

[14] WJ. Gauderman, E. Avol, F. Lurmann, et al. Childhood asthma and exposure to traffic and nitrogen dioxide. Epidemiology;16:737–43, 2005.

[15] RJ. Carroll, D. Ruppert, LA. Stefanski, et al. Measurement error in nonlinear models. London, United Kingdom: DC.Thomas, D. Stram, J. Dwyer, et al. Exposure measurement error: influence on exposure-disease relationships and methods of correction. Annu Rev Public Health;14:69–93, 1993.

[16] WR. Gilks, S. Richardson, DJ. Spiegelhalter. Markov chain Monte Carlo in practice. London, United Kingdom: Chapman and Hall, 1996.

[17] DW. Dockery, J. Cunningham, AI. Damokosh, et al. Health effects of acid aerosols on North American children: respira-tory symptoms. Environ Health Perspect;104:50, 1996.

[18] WJ. Gauderman, GF. Gilliland, H.Vora, et al. Association be-tween air pollution and lung function growth in southern California children: results from a second cohort. Am J Respir Crit Care Med;166:76–84, 2002.

[19] P. Benson. CALINE4—a dispersion model for predicting air pollution concentration near roadways. Sacramento, CA: Of-fice of

[20] MC. Roorda-Knape, NA. Janssen, J. Hartog, et al. Traffic related air pollution in city districts near motorways. Sci Total Environ;253:339–41, 1999.

[21] W. Fuller. Measurement error models. New York, NY: John Wiley & Sons, 1987.

[22] D. Spiegelhalter, A. Thomas, N. Best. WinBUGS version 1.4 user manual. Cambridge, MA: MRC Biostatistics Unit, 2003.

[23] A.Gelman, JB. Carlin, HS. Stern, et al. Bayesian data analysis. Boca Raton, FL: Chapman & Hall/CRC, 1995.

[24] DB. Rubin. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley & Sons, 1987.

[25] JL.Schafer. Analysis of incomplete multivariate data. London, United Kingdom: Chapman & Hall, 1997.

[26] SAS language reference: dictionary, version 8. Cary, NC: SAS Institute, Inc, 2004.