Monte Carlo Adaptive Technique for Sensitivity Analysis of a Large-scale Air Pollution Model

Ivan Dimov[1,2] and Rayna Georgieva[2]

[1] ACET, The University of Reading
Whiteknights, PO Box 225, Reading, RG6 6AY, UK
[2] IPP, Bulgarian Academy of Sciences
Acad. G. Bonchev 25 A, 1113 Sofia, Bulgaria
i.t.dimov@reading.ac.uk, rayna@parallel.bas.bg

Abstract. Variance-based sensitivity analysis has been performed for a study of input parameters contribution into output variability of a large-scale air pollution model - the Unified Danish Eulerian Model. The problem of computing of numerical indicators of sensitivity - Sobol' global sensitivity indices leads to multidimensional integration. Plain and Adaptive Monte Carlo techniques for numerical integration have been analysed and applied. Numerical results for sensitivity of pollutants concentrations to chemical rates variability are presented.

1 Introduction

Sensitivity **a**nalysis (SA) is a study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input [6]. There are several available sensitivity analysis techniques [6]. Variance-based methods deliver results that are independent to the models behaviors: linearity, monotonicity and additivity of the relationship between input factor and model output sensitivity measures.

The aim of our research is to develop an Adaptive Monte Carlo (MC) algorithm for evaluating Sobol' global sensitivity indices increasing the reliability of the results by reducing the variance of the corresponding Monte Carlo estimator and applying adaptive concept to numerical integration of functions with local difficulties. It gives a possibility for a robust SA - a comprehensive study the influence of variations of the chemical rates on the model results in a particular case. This helps to identify the most influential parameters and mechanisms and to improve the accuracy or insignificant parameters. It also helps to simplify the model by ignoring them after careful complementary analysis of relations between parameters.

2 Background studies

The investigations and the numerical results reported in this paper have been obtained by using a large-scale mathematical model called **Unified Danish**

Eulerian Model (UNI-DEM) [7, 8]. This model simulates the transport of air pollutants and has been developed by Dr. Z. Zlatev and his collaborators at the Danish National Environmental Research Institute (http://www2.dmu.dk/AtmosphericEnvironment/DEM/). Both non-linearity and stiffness of the equations are mainly introduced by the chemistry (CBM-4 chemical scheme) [8]. Thus, the motivation to choose UNI-DEM is that it is one of the models of atmospheric chemistry, where the chemical processes are taken into account in a very accurate way. Our main interest is to find out how changes in the input parameters of the model influence the model output. We consider the chemical rate constants to be input parameters and the concentrations of pollutants to be output parameters. In the context of this paper, the term "constants" means variables with normal distribution (established experimentally) with mean 1.0.

2.1 Sobol' Global Sensitivity Indices Concept

It is assumed that the mathematical model can be presented as a model function

$$\mathrm{u} = f(\mathrm{x}), \quad \text{where} \quad \mathrm{x} = (x_1, x_2, \ldots, x_d) \in U^d \equiv [0;1]^d \tag{1}$$

is a vector of input independent parameters with a joint **p**robability **d**ensity **f**unction (p.d.f.) $p(\mathrm{x}) = p(x_1, \ldots, x_d)$. The **t**otal **s**ensitivity **i**ndex (TSI) of input parameter $x_i, \;\; i \in \{1, \ldots, d\}$ is defined in the following way [3]: $S_{x_i}^{tot} = S_i + \sum\limits_{l_1 \neq i} S_{il_1} + \sum\limits_{l_1, l_2 \neq i, l_1 < l_2} S_{il_1l_2} + \ldots + S_{il_1 \ldots l_{d-1}}$, where S_i is called *the main effect (first-order sensitivity index)* of x_i and $S_{il_1 \ldots l_{j-1}}$ is the $j^{\text{-th}}$ order sensitivity index (respectively *two-way interactions* for $j = 2$, and so on) for parameter x_i $(2 \leq j \leq d)$.

The variance-based Sobol' method [3] uses the sensitivity measures (indices) and takes into account interaction effects between inputs. An important advantage of this method is that it allows to compute not only the first-order indices, but also indices of a higher-order in a way similar to the computation of the main effects, the total sensitivity index can be calculated with just one Monte Carlo integral per factor. The computational cost of estimating all first-order $(m = 1)$ and total sensitivity indices via Sobol' approach is proportional to dN where N is the sample size and d is the number of input parameters (see [2]). The method is based on a decomposition of an integrable model function f in the d-dimensional factor space into terms of increasing dimensionality:

$$f(\mathrm{x}) = f_0 + \sum_{\nu=1}^{d} \sum_{l_1 < \ldots < l_\nu} f_{l_1 \ldots l_\nu}(x_{l_1}, x_{l_2}, \ldots, x_{l_\nu}), \quad f_0 = \int f(\mathrm{x}) \mathrm{dx} \tag{2}$$

where f_0 is a constant. The representation (2) is unique (called ANOVA - representation of the model function $f(\mathrm{x})$ [4]) if $\int_0^1 f_{l_1 \ldots l_\nu}(x_{l_1}, x_{l_2}, \ldots, x_{l_\nu}) \mathrm{d}x_{l_k} = 0, \quad 1 \leq k \leq \nu, \quad \nu = 1, \ldots, d$. The quantities $\mathbf{D} = \int_{U^d} f^2(\mathrm{x}) \mathrm{dx} - f_0^2, \;\; \mathbf{D}_{l_1 \ \ldots \ l_\nu} = \int f_{l_1 \ \ldots \ l_\nu}^2 \underset{.}{\mathrm{x}}_{l_1} \ldots \underset{.}{\mathrm{x}}_{l_\nu}$ are called variances (total and partial variances, respectively), where $f(\mathrm{x})$ is a square integrable function. Therefore, the total variance of the model output is partitioned into partial variances [3] in the analogous way as the model function, that is the ANOVA-decomposition: $\mathbf{D} = \sum_{\nu=1}^{d} \sum_{l_1<\ldots<l_\nu} \mathbf{D}_{l_1 \ldots l_\nu}$. Based on the above assumptions about the model function and the output variance, the following quantities

$$S_{l_1 \ \ldots \ l_\nu} = \frac{\mathbf{D}_{l_1 \ \ldots \ l_\nu}}{\mathbf{D}}, \quad \nu \in \{1, \ldots, d\} \tag{3}$$

are called Sobol' global sensitivity indices [3, 4].

The results discussed above make clear that the mathematical treatment of the problem of providing global sensitivity analysis consists in evaluating total sensitivity indices and in particular Sobol' global sensitivity indices (3) of corresponding order. This leads to computing of multidimensional integrals $I = \int_{\Omega} g(\mathrm{x})p(\mathrm{x})\,\mathrm{dx}, \ \ \Omega \subset \mathbf{R}^d$, where $g(\mathrm{x})$ is a square integrable function in Ω and $p(\mathrm{x}) \geq 0$ is a p.d.f., such that $\int_{\Omega} p(\mathrm{x})\,\mathrm{dx} = 1$.

The procedure for computation of global sensitivity indices is based on the following representation of the variance $\mathbf{D}_{\mathrm{y}}$: $\mathbf{D}_{\mathrm{y}} = \int \ f(\mathrm{x}) \ f(\mathrm{y}, \mathrm{z}')\mathrm{dxdz}' - f_0^2$ (see [4]), where $\mathrm{y} = (x_{k_1}, \ldots, x_{k_m}), \quad 1 \leq k_1 < \ldots < k_m \leq d$, is an arbitrary set of m variables $(1 \leq m \leq d-1)$ and z be the set of $d-m$ complementary variables, i.e. $\mathrm{x} = (\mathrm{y}, \mathrm{z})$. Let $K = (k_1, \ldots, k_m)$ and the complement of the subset K in the set of all parameter indices is denoted by $\bar{K}$. The last equality allows to construct a Monte Carlo algorithm for evaluating $f_0, \mathbf{D}$ and $\mathbf{D}_{\mathrm{y}}$, where $\xi = (\eta, \zeta)$:

$$\frac{1}{N} \sum_{j=1}^{N} f(\xi_j) \xrightarrow{P} f_0, \qquad \frac{1}{N} \sum_{j=1}^{N} f(\xi_j) \ f(\eta_j, \zeta_j') \xrightarrow{P} \mathbf{D}_{\mathrm{y}} + f_0^2,$$

$$\frac{1}{N} \sum_{j=1}^{N} f^2(\xi_j) \xrightarrow{P} \mathbf{D} + f_0^2, \qquad \frac{1}{N} \sum_{j=1}^{N} f(\xi_j) \ f(\eta_j', \zeta_j) \xrightarrow{P} \mathbf{D}_{\mathrm{z}} + f_0^2.$$

2.2 Monte Carlo Approach for Small Sensitivity Indices

The standard Monte Carlo algorithm for estimating global sensitivity indices, proposed in [3], is spoilt by loss of accuracy when $\mathbf{D}_y << f_0^2$, i.e. in the case of small (in values) sensitivity indices. That is why here we have applied one of the approaches for evaluating small sensitivity indices - the so called combined approach [2]. The concept of the approach consists in replacement of the original integrand (the mathematical model function) with a function of the following type $\varphi(\mathrm{x}) = f(\mathrm{x}) - c$, where $c \sim f_0$. The following estimator for variances has been proposed for this approach:

$$\mathbf{D}_{\mathrm{y}} = \int \varphi(\mathrm{x}) \ [\varphi(\mathrm{y}, \mathrm{z}')\mathrm{dxdz}' - \varphi(\mathrm{x}')]\mathrm{dxdx}', \ \mathbf{D} = \int \varphi(\mathrm{x})[\varphi(\mathrm{x}) - \varphi(\mathrm{x}')]\mathrm{dxdx}'.$$

3 Description of the Algorithms

Two Monte Carlo algorithms have been applied: Plain and Adaptive. **Plain (Crude) Monte Carlo** is the simplest possible MC approach for solving multidimensional integrals [1]. Let us consider the problem of the approximate computation of the integral $I = \int_{\Omega} g(\mathrm{x})p(\mathrm{x})\mathrm{dx}$. Let ξ be a random point with a p.d.f. $p(\mathrm{x})$. Introducing the random variable $\theta = f(\xi)$ such that $\mathbf{E}\theta = \int_{\Omega} g(\mathrm{x})p(\mathrm{x})\mathrm{dx}$. Let the random points $\xi_1, \xi_2, \dots, \xi_N$ be independent realizations of the random point ξ with p.d.f. $p(\mathrm{x})$ and $\theta_1 = f(\xi_1), \dots, \theta_N = f(\xi_N)$. Then an approximate value of I is $\overline{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} \theta_i$. The last equation defines the Plain Monte Carlo algorithm.

There are various **Adaptive Monte Carlo algorithms** depending on the technique of adaptation [1]. Our Adaptive algorithm uses a posteriori information about the variance. The idea of the algorithm consists in the following: the domain of integration Ω is separated initially into subdomains with identical volume. The corresponding interval on every dimension coordinate is partitioned into M subintervals, i.e. $\Omega = \sum_j \Omega_j,\ \ j = 1,\, M^d$. Denote by p_j and I_{Ω_j} the following expressions: $p_j = \int_{\Omega_j} p(\mathrm{x})\,\mathrm{dx}$ and $I_{\Omega_j} = \int_{\Omega_j} f(\mathrm{x})p(\mathrm{x})\,\mathrm{dx}$. Consider now a random point $\xi^{(j)} \in \Omega_j$ with a density function $p(\mathrm{x})/p_j$. In this case $I_{\Omega_j} = \mathbf{E}\left[\frac{p_j}{N}\sum_{i=1}^{N} f(\xi_i^{(j)})\right] = \mathbf{E}\theta_N$.

The algorithm starts with a relatively small number M which is given as input data. For every subdomain the integral I_{Ω_j} and the variance are evaluated. Then the variance is compared with a preliminary given value. The obtained information is used for the next refinement of the domain and for increasing the density of the random points. The subdomain with the largest variance is divided onto 2^d new subdomains. The algorithm stops when the variance at all obtained after division subdomains satisfies the preliminary given accuracy ε (or when a given maximum value of number of levels or subdomains where the stop criterion is not satisfied has been reached).

4 Analysis of Numerical Results and Discussion

The first stage of computations includes a generation of input data for our procedure using the UNI-DEM. The model runs have been done for the chemical rates variations with a fixed set of perturbation factors $\alpha = \{\alpha_i\}, i = 1, \dots, d$, where every α_i corresponds to a chemical rate among the set of 69 time-dependent chemical reactions and 47 constant chemical reactions, and d is the total number of chemical reactions taken into account in the numerical experiments. The generated data is ratios of the following type $r_s(\alpha) = \dfrac{c_s^{\alpha}(a_s^{i_{max}}, b_s^{j_{max}})}{c_s^{max}}, \quad \alpha_i \in \{0.1, 0.2, \dots, 2.0\}$, where the lower index s corresponds to the chemical species (pollutants). The denominator $c_s^{max} = c_s^{max}(a_s^{i_{max}}, b_s^{j_{max}})$ is the maximum mean value of the concentration of chemical species s obtained for $\alpha = (1, \dots, 1)$, i.e. without any perturbations, $a_s^{i_{max}}$ and $b_s^{j_{max}}$ are the coordinates of the point, where this maximum has been reached, and i_{max}, j_{max} are the mesh indices of this point. The nominator represents the values of the concentrations of the corresponding pollutant for a given set of values of the perturbation parameters $\alpha_i \in \{0.1, \dots, 2.0\}$, computed at the point $(a_s^{i_{max}}, b_s^{j_{max}})$. Thus we consider a set of pollutant concentrations normalized according to the maximum mean value of the concentration of the corresponding chemical species. We also study numerically how different chemical rate reactions influence concentrations of a given pollutant. An example of how chemical rate reaction of three different reactions (## 3, 6 and 22 of CBM-4 chemical scheme) influence ozone concentrations is presented on Figure 1. One can see that in this particular case the influence of reactions ## 3 and 22 is significant and the influence of reaction 6 is almost negligible.

Sensitivity Analysis computations consists of two steps: approximation and computing of Sobol' global sensitivity indices. As a result of computations with the use of UNI-DEM we obtain mesh functions in a form of tables of the values of the model function. The first step is to represent the model as a continuous function (1). To do that we use approximation by polynomials of third and forth degree, where $p_s(\mathrm{x})$ is the polynomial approximating the mesh function that corresponds to the s-th chemical species. The approximation domain $\Omega = [0.1; 2.0]^3$ has been chosen a bit wider ranging than the integration domain $\Omega = [0.6; 1.4]^3$ in order to present more precisely the mesh model function. The squared 2-vector norm of the residual defined as $\| p_s - r_s \|_2^2 = \sum_{l=1}^{n} [p_s(\mathrm{x}_l) - r_s(\mathrm{x}_l)]^2, \mathrm{x}_l \in [a; b]^3$ in the case of a polynomial of 4-th degree in three variables was $\| p_s - r_s \|_2^2 = 0.022$ for $\mathrm{x}_l \in [0.1; 2.0]^3$ and $\| p_s - r_s \|_2^2 = 0.00022$ for $\mathrm{x}_l \in [0.6; 1.4]^3$ in our numerical experiments. If one is not happy with the accuracy of polynomial approximation other tools should be used. So, we consider polynomials of 4-th degree in three variables as a case study which is completely satisfying us at this stage.

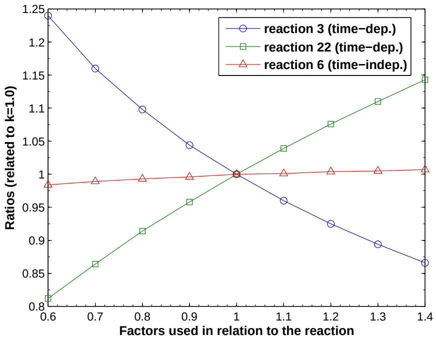

Fig. 1. Sensitivity of ozone concentrations to changes of chemical rates.

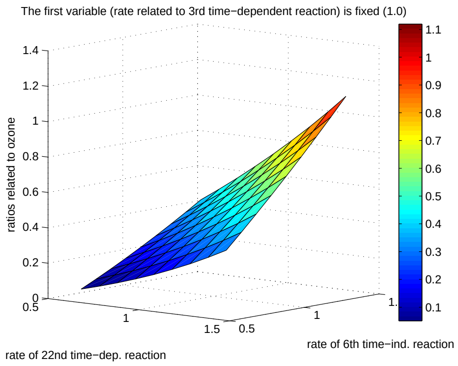

Fig. 2. Model function for $x_1 = 1$.

The approximate function is smooth but it has a single peak at one of the corners of the domain. A section of the model function graphics (the first variable is fixed to 1.0) is presented on Figure 2. The adaptive approach seems to be promising for functions like this. Adaptive Monte Carlo algorithm (see Section 3) has been applied to the problem of numerical integration. The results have

Table 1. First-order and total sensitivity indices of input parameters estimated using different approaches of sensitivity analysis applying Plain Monte Carlo algorithm.

approach quantity	Standard (Sobol')			Combined		
	N	Est. value	Rel. error	N	Est. value	Rel. error
g_0	10^4	0.5155	8e-05	10^4	0.2525	0.0002
	10^6	0.5156	3e-05	10^6	0.2526	7e-05
	10^7	0.5155	2e-05	10^7	0.2526	5e-05
D	10^4	0.2635	0.0002	10^4	0.0052	0.0081
	10^6	0.2636	6e-05	10^6	0.0052	0.0009
	10^7	0.2636	5e-05	10^7	0.0052	0.0012
S_1	10^4	0.2657	0.0127	10^4	0.5349	0.0048
	10^6	0.2654	0.0116	10^6	0.5337	0.0025
	10^7	0.2653	0.0114	10^7	0.5325	0.0003
S_3	10^4	0.2525	0.0013	10^4	0.0012	0.4060
	10^6	0.2521	0.0002	10^6	0.0019	0.0333
	10^7	0.2521	0.0002	10^7	0.0019	0.0213
$S_{x_1}^{tot}$	10^4	0.4183	0.0002	10^4	0.5389	0.0021
	10^6	0.4185	0.0003	10^6	0.5392	0.0026
	10^7	0.4184	0.0001	10^7	0.5380	0.0003
$S_{x_3}^{tot}$	10^4	0.4017	0.0002	10^4	0.0009	0.6078
	10^6	0.4018	5e-06	10^6	0.0022	0.0354
	10^7	0.4018	8e-06	10^7	0.0022	0.0192

been compared with Plain MC (see Section 3). One of the best available random number generators, SIMD-oriented Fast Mersenne Twister (SFMT) [5] 128-bit pseudorandom number generator of period $2^{19937} - 1$ has been used to generate the required random points.

Results for some first-order and total sensitivity indices obtained by the Plain Monte Carlo algorithm are presented in Table 1. Two approaches for sensitivity indices have been applied - standard (Sobol', 2001) and combined approach. For the implementation of the combined approach a Monte Carlo estimate of f_0 has been used - $c = 0.51365$. The following notation is used in the tables: ε is the desired estimate of standard deviation, $\#_{sub}$ is the number of subdomains after domain division, N_{sub} is the number of samples in each subdomain, N is the number of samples in the domain of integration (for the Plain algorithm), D is the variance; g_0 is the integral over the domain $[0.6; 1.4]^3$, where the integrand for the standard approach is the model function, $f(x)$, and for the combined approach $f(x) - c$ respectively. Relative error is the absolute error divided by the exact value. Each estimated value is obtained after 10 algorithm runs. The exact values are know for our special case of mesh function approximation.

One of the advantages of Sobol' type approaches has been applied in the implementation of the Plain Monte Carlo algorithm - the possibility to compute first-order and total sensitivity indices of a given input parameter using only one Monte Carlo integral and two independent multidimensional sequences of random numbers. The results obtained confirm the expected effect of decrease of the relative error with the increase of the number of samples. On the other

Table 2. First-order and total sensitivity indices of input parameters estimated using combined approach applying Plain and Adaptive Monte Carlo algorithms.

Estimated	Plain			Adaptive				
quantity	N	Est. value	Rel. error	$\#_{sub}$	N_{sub}	ε	Est. value	Rel. error
g_0	192	0.2516	0.0038	-	-	-	-	-
	7200	0.2524	0.0005	-	-	-	-	-
	32000	0.2524	0.0005	-	-	-	-	-
D	192	0.0055	0.0493	64	3	0.5	0.0056	0.0725
	7200	0.0052	0.0130	180	40	0.0165	0.0051	0.0333
	32000	0.0052	0.0034	64	500	0.1	0.0052	0.0003
S_1	192	0.6502	0.2214	64	3	0.5	0.5072	0.0473
	7200	0.5299	0.0046	180	40	0.0165	0.5307	0.0031
	32000	0.5326	0.0004	64	500	0.1	0.5323	0.0001
S_3	192	0.0055	1.7695	64	3	0.5	0.0016	0.1790
	7200	0.0009	0.5367	64	500	0.1	0.0027	0.3463
	32000	0.0013	0.3250	64	10^4	0.1	0.0019	0.0581
$S_{x_1}^{tot}$	192	0.5875	0.0923	64	3	0.5	0.5108	0.0503
	7200	0.5346	0.0061	180	40	0.0165	0.5345	0.0061
	32000	0.5368	0.0020	64	500	0.1	0.5376	0.0004
$S_{x_3}^{tot}$	192	0.0004	1.1693	64	3	0.5	0.0047	1.1013
	7200	0.0006	0.7529	180	40	0.0165	0.0021	0.0895
	32000	0.0018	0.2094	64	500	0.1	0.0022	0.0153

hand, the order of relative error decreases for the values of sensitivity indices in comparison with g_0 and variation D for both approaches. These quantities are presented by only one (g_0) or two (D) integrals while each sensitivity index (first-order or total) is presented by a ratio of integrals estimated be the Plain Monte Carlo algorithm that leads to an accumulation of errors. The value of variation for the standard approach is much smaller than the value of variation for the combined approach and the division into that relatively small quantity leads to larger relative errors for total sensitivity indices using the combined approach. Nevertheless, the results for total sensitivity indices obtained by the combined approach are more reliable - the values of total effects are fully consistent with the expected tendencies according to Figure 1.

A comparison between computed first-order and total sensitivity indices obtained by the combined approach using Plain and Adaptive Monte Carlo algorithms is given in Table 2. The concepts of the combined approach and the developed adaptive approach require numerical integration over 6-dimensional domain, i.e. twice as large as the dimension of the model function. That is why g_0 has not been computed in this case. Since the total number of estimated quantities is seven - variance and main and total effects of three input parameters - the adaptive procedure applied to all of them would be inefficient according its computational cost. The criterion for achieving the desired accuracy in computing variance has been adopted as a common criterion in computing other quantities because all main and total effects depend on the variance. In contrast of that, two independent random sequences in 3-dimensional domain $[0.6; 1.4]^3$ have been used for implementation of the Plain Monte Carlo algorithm.

The number of samples for both Monte Carlo techniques has been chosen following the requirement for consistency of obtained results, i.e. the number of samples for Plain algorithm is a multiplication of average number of subdomains (from several runs) and number of samples in each subdomain. It has been observed that the computational times using two Monte Carlo approaches for numerical integration (Plain and Adaptive) to estimate the unknown quantities are comparable. For example, $t = 0.073s$ for the Plain ($N = 32\ 000$) and $t = 0.078s$ for the Adaptive. Thus, the conclusions about efficiency of the applied algorithms in computing the desired quantities can be made by comparing orders of estimated errors.

The Adaptive algorithm has an advantage over the Plain algorithm for a fixed number of samples that confirms reducing variance effect of the applied adaptive technique. Moreover, the approximative values of quantities are sufficiently close to exact values even for the smallest chosen number of samples.

Aknowledgement

The research reported in this paper is partly supported by NATO Science under grant CLG.982641, as well as by the Bulgarian NSF Grants DO 02-215/2008 and DO 02-115/2008. The authors thank Z. Zlatev and Tz. Ostromsky for providing necessary data for our analysis and for stimulating discussions.

References

1. I. Dimov, *Monte Carlo Methods for Applied Scientists* (World Scientific, Singapore, 2008).
2. A. Saltelli, Making best use of model valuations to compute sensitivity indices, *Computer Physics Communications* **145** (2002) 280–297.
3. I.M. Sobol', Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment* **1** (1993) 407–414.
4. I.M. Sobol', Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Mathematics and Computers in Simulation*, **55 (1-3)** (2001) 271–280.
5. Mutsuo Saito, Makoto Matsumoto, SIMD-oriented Fast Mersenne Twister: a 128-bit Pseudorandom Number Generator, *Monte Carlo and Quasi-Monte Carlo Methods 2006*, Springer (2008) 607–622.
6. Saltelli A., Tarantola S., Campolongo, F. and Ratto, M.,*Sensitivity Anal- ysis in Practice. A Guide to Assessing Scientific Models* , John Wiley & Sons publishers (2004).
7. Z. Zlatev, *Computer treatment of large air pollution models*, KLUWER Academic Publishers (Dorsrecht-Boston-London, 1995).
8. Z. Zlatev, I. Dimov, *Computational and Numerical Challenges in Environmental Modelling* (Elsevier, Amsterdam, 2006).