Research Journal of Chemical Sciences ISSN 2231-606X
Vol. **4(7),** 24-29, July **(2014)** *Res. J. Chem. Sci.*

Comparative Evaluation of Multiple Linear Regression and Support vector Machine aided Linear and Non-linear QSAR Models

Joshi Shobha[1], Sharma Sonal[2], Yadav Mukesh[3*]
[1]Govt. Shaheed Bhagirath Silawat College, Depalpur, Indore, MP, INDIA
[2]Dept. of Chemistry, Govt. Holkar Science College, Indore, MP, INDIA
[3]Dept. of Pharmaceutical Chemistry, Softvision College, Indore, MP, INDIA

Available online at: www.isca.in, www.isca.me
Received 16th May 2014, revised 17th June 2014, accepted 14th July 2014

Abstract

Type 2 diabetes still remains a major challenge to human health management. Protein tyrosine phosphate 1B has been continuously explored for its therapeutic potential to treat type 2 diabetes as it is linked with negative regulation of insulin signal transduction. QSAR studies were performed on derivatives of 2-arylsulphonylaminobenzothiazoles. MLR aided linear and SVM aided linear and non-linear models were obtained which were further evaluated to identify descriptors revealing underlying structure-activity relationship. QSAR models were validated through a series of validation techniques like Y-randomization and descriptor sensitivity in addition to internal validation parameters. Information content index (IC1) of neighbourhood symmetry of order-1 has been found to be a key molecular descriptor participating and regulating structure–activity relationship of 2-arylsulphonylaminobenzothiazoles derivatives. Geary auto correlated atomic masses and polarizability are also actively correlated to biological response of tyrosine phosphate 1B inhibitors.

Keywords: Type 2 diabetes, tyrosine phosphate 1B inhibitors, Linear and non-linear QSAR models, MLR, SVM.

Introduction

Mellitus, the type 2 diabetes is chronic and progressive disease of metabolic disorder. Obesity and insulin resistance are the very common risk factors of developing type 2 diabetes mellitus. Diabetes involves the high level of glucose in blood plasma. Many people with type 2 diabetes mellitus have hypertension and high level of cholesterol. All of these factors can cause the long term complication such as neuropathy, retinopathy, nephropathy and cardiovascular disorder *etc*[1,2].

Protein tyrosine phosphate 1B was separated through the process of distillation from human placental tissue in 1988 and crystallized in 1994 [3]. Protein tyrosine phosphate 1B is the most flourishing molecular level rational therapeutic target in the efficacious direction of treatment of type 2 diabetes mellitus. Cytosolic nonreceptor protein tyrosine phosphatase, PTP1B, is key factor in the negative regulation of insulin signal transduction[4,5]. PTP1B inhibitors block the PTP1B mediated negative insulin signal transduction and leads to stimulation of insulin activity. Therefore, they can be considered as the most fascinating target for type 2 diabetes mellits[6,7]. The inhibitors of PTP1B are classified into four categories: difluoromethylene phosphates, 2-carbomethoxybenzoicacid, 2-oxalyl amino benzoic and hydrophobic compound[4].

QSAR methods attempt to find out the relationship between end point (Biological activity) and chemical structures, which allows the prediction of potency of drug[8-10]. Machine learning is the field of artificial intelligence associated with study of computer algorithm that improves on its own through experience[11]. There are few machine learning approaches such as ANN[12], SVM[13], Decision Tree[14] and Bays Classifier[15] which is used in QSAR modelling while multiple linear regression is most extensively used method to construct QSAR models[16].

Methodology

Dataset of twenty seven (27) derivatives of 2-arylsulphonylaminobenzothiazole were taken from literature[17] for QSAR study. The 3D structures of molecules were drawn by software Marvin Sketch 5.1.5 (developed by Chemaxon Ltd.)[18]. Structural details and experimental biological activity are reported in table-1.

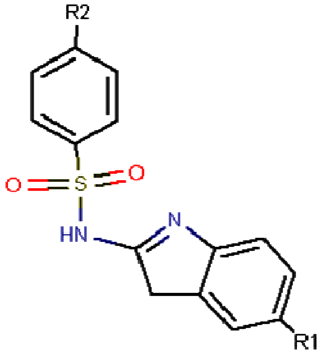

Figure-1
Structure of 2-arylsulphonylaminobenzothiazole and scheme of derivatives

Table-1
Structural details and experimental biological activity for derivatives 1-27

Compound	R_1	R_2	Antiobesity effect(*Pa/Pi*)
1	$-CH_3$	-H	0.892
2	$-CH_3$	$-CH_3$	0.897
3	$-CH_3$	$-OCH_3$	0.853
4	$-CH_3$	$-NO_2$	0.835
5	$-CH_3$	$-NHCOCH_3$	0.819
6	$-CH_3$	Cl	0.884
7	$-OCH_3$	H	0.889
8	$-OCH_3$	$-CH_3$	0.863
9	$-OCH_3$	$-OCH_3$	0.894
10	$-OCH_3$	$-NO_2$	0.833
11	$-OCH_3$	$-NHCOCH_3$	0.809
12	$-OCH_3$	Cl	0.882
13	$-OCH_2CH_3$	H	0.880
14	$-OCH_2CH_3$	$-CH_3$	0.858
15	$-OCH_2CH_3$	$-OCH_3$	0.858
16	$-OCH_2CH_3$	$-NO_2$	0.817
17	$-OCH_2CH_3$	$-NHCOCH_3$	0.813
18	$-OCH_2CH_3$	Cl	0.872
19	$-NO_2$	H	0.888
20	$-NO_2$	$-CH_3$	0.860
21	$-NO_2$	$-OCH_3$	0.845
22	$-NO_2$	$-NO_2$	0.893
23	$-NO_2$	$-NHCOCH_3$	0.809
24	$-NO_2$	Cl	0.880
25	$-NO_2$	F	0.861
26	-F	$-NO_2$	0.836
27	-Cl	$-NO_2$	0.881

Descriptors for each derivative of corresponding compound were computed by E-Dragon software[19]. A large pool of significant descriptors was calculated for each molecule. Highly correlated descriptors and descriptors, having constant values, missing values or zero value were removed in pruning. In QSAR study of 2-arylsulphonylaminobenzothiazole derivatives molecular descriptors subset was obtained by forward selection method[20]. Linear QSAR models were developed, using the most simple and popular method multiple linear regression and support vector machine aided linear method while non-linear QSAR models were developed using Gaussian kernel function aided support vector machine[21]. Support vector machine classifies data by constructing best hyper plane by applying kernel trick to separate molecules into two classes[22]. There are three types of kernel functions are available linear, Gaussian and polynomial. Robustness of QSAR models, obtained after linear and non-linear regressions was evaluated by using leave one out internal cross validation (R^2_{CV}) and predictive error sum of square PRESS.

To protect against chance correlation Y-randomization method was performed[23,24]. Descriptor sensitivity analysis was performed to identify the most sensitive descriptor[25].

Results and Discussion

Multiple linear regressions: In step-wise multiple linear regressions we have obtained three significant QSAR models. Tri-variable model was selected as most significant model and it is represented below equation-1.

$$Pa/pI= 2.1464- 0.7575\ (0.8735)\ [GATS3m] - 0.1624\ (0.0093)\ [IC1] + 0.0056\ (0.0014)\ [RDF105m] \quad (1)$$

$N = 27$, $R^2 = 0.9393$, $R^2A = 0.9303$, $F = 116.63$, $S.E. = 0.014$, $R^2_{CV} = 0.9117$, $S_{PRESS} = 0.1905$, $RSS = 0.0196$

Herein, N = number of compounds, R^2 = coefficient of determination, R^2A is adjusted R^2, F = Fisher's statistics and S.E. = standard error. R^2_{CV} = leave one out (LOO) cross validation parameter and S_{PRESS} = standard deviation based on predictive error sum of square, RSS is residual sum of square. Among three QSAR models statistical significant value of R^2_{CV} (0.9117), lowest value of PRESS (0.1905), highest value of RSS (0.0196) prove tri-variable model as the most predictive and statistically fit model.

SVM regression: We have used linear kernel function for SVM aided linear regression and Gaussian kernel function for SVM aided non-linear regression at fixed value of cost function C (100), ε- insensitive loss function (0.1) and sigma (0.1). Statistical parameters R^2, S.E., and validation parameters $R^2_{CV,}S_{PRESS}$ and RSS for MLR and SVM aided linear and non-linear regressions are summarised in table-2.

Descriptor IC1 (neighbourhood symmetry of 1-order) was selected for model building by MLR and SVM aided linear and non-linear regression methods. IC1 belongs to information content descriptors. Descriptor SIC3 (Structural information content -neighbourhood symmetry of order-3) was selected in SVM aided linear models. Descriptor RDF105u (Radial distribution function-10.5/unweighted) belongs to radial distribution function descriptors and was selected in MLR QSAR models. Descriptor GATS3m (Geary autocorrelation – lag3 weighted by atomic masses) was selected in MLR aided linear and SVM aided non-linear QSAR models. Descriptor GATS3m belongs to 2D Autocorrelation indices. GATS5p (Geary autocorrelation – lag5 weighted by atomic polarizability) belongs to 2D Autocorrelation indices was selected in SVM aided non linear modelling. Descriptor Mor07v (signal 7 weighted by atomic Vander walls volume) belongs to 3D-MORSE descriptors. It has been found to play significant role in SVM aided linear QSAR models. All descriptors selected in MLR and SVM aided linear and non-linear models are listed in table-3.

Validation: Calculated biological activity *Pa/Pi* in MLR and SVM aided linear and non-linear QSAR models show good correlation with experimental activity and show robustness of models table-4 and figure-2.

Table-2
Statistical fitness parameters and validation parameters used in MLR and SVM aided linear and non-linear QSAR models

	MLR			**SVM aided linear QSAR**			**SVM aided non-linear**		
Model	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
R^2	0.731	0.892	0.939	0.707	0.865	0.925	0.770	0.872	0.954
SE	0.030	0.018	0.014	0.033	0.024	0.017	0.031	0.031	0.031
RSS	0.006	0.020	0.020	0.006	0.003	0.002	0.005	0.005	0.005
R^2_{cv}	0.689	0.866	0.912	0.686	0.844	0.910	0.746	0.852	0.915
PRESS	0.192	0.191	0.191	0.192	0.192	0.192	0.192	0.192	0.192

Table-3
Descriptors used in MLR and SVM aided linear and nonlinear modelling

Descriptors	**MLR**
IC1	Information content index(neighborhood symmetry of 1-order)
GATS3m	Geary autocorrelation-lags3/weighted by atomic masses
RDF105m	Radial Distribution Function - 10.5 / weighted by atomic masses
Descriptors	**SVM aided linear QSAR study**
Mor07v	3-D mores signal 7/weighted by atomic Vander Waals volumes
IC1	Information content index(neighborhood symmetry of 1-order)
SIC3	structural information content (neighborhood symmetry of 3-order)
Descriptors	**SVM aided non-linear QSAR study**
IC1	Information content index(neighborhood symmetry of 1-order)
GATS3m	Geary autocorrelation-lags3/weighted by atomic masses
GATS5p	Geary autocorrelation-lags5/weighted by atomic polarizability

Table-4
Experimental activity and calculated activity of 2-arylsulphonyl amino benzothiazole derivatives for MLR and SVM aided linear and nonlinear modelling.

Compound	Experimental (Antiobesity)	MLR	SVM aided linear	SVM aided non-linear
		Calculated (antiobesity)	Calculated (antiobesity)	Calculated (antiobesity)
1	0.892	0.893	0.894	0.895
2	0.897	0.894	0.903	0.892
3	0.853	0.863	0.852	0.862
4	0.835	0.841	0.839	0.840
5	0.819	0.828	0.826	0.826
6	0.884	0.883	0.885	0.887
7	0.889	0.883	0.890	0.888
8	0.863	0.866	0.866	0.864
9	0.894	0.880	0.894	0.874
10	0.833	0.829	0.836	0.827
11	0.809	0.811	0.816	0.808
12	0.882	0.875	0.882	0.881
13	0.880	0.869	0.873	0.874
14	0.858	0.860	0.854	0.857
15	0.858	0.866	0.856	0.857
16	0.817	0.821	0.823	0.819
17	0.813	0.810	0.805	0.805
18	0.872	0.866	0.867	0.873
19	0.888	0.902	0.888	0.891
20	0.860	0.854	0.843	0.856
21	0.845	0.842	0.828	0.851
22	0.893	0.890	0.878	0.878
23	0.809	0.811	0.809	0.809
24	0.880	0.894	0.877	0.880
25	0.861	0.852	0.847	0.860
26	0.836	0.836	0.847	0.839
27	0.881	0.878	0.865	0.879

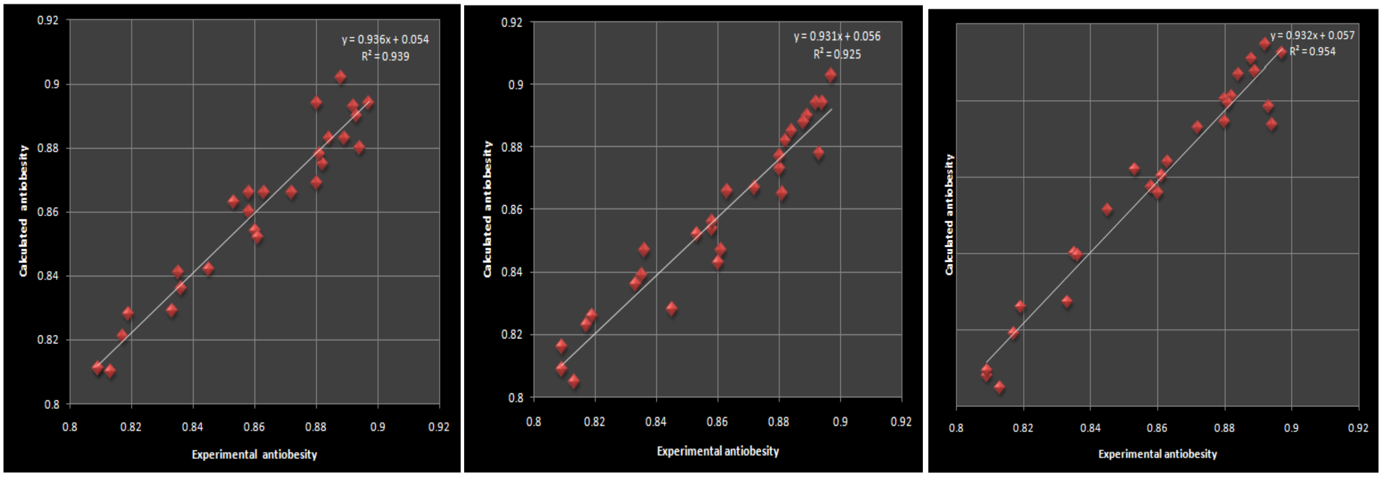

Figure-2
Correlation of experimental vs. calculated antiobesity of (a) MLR aided tri-variable model, (b) SVM aided linear tri-variable model, (c) SVM aided non-linear tri-variable model

Figure-2a represents linear relationship of the experimental activity with calculated activity for MLR aided tri-variable model. Prediction of biological activity is found to be satisfactory with R^2 value (0.9393) while figure-2b represents linear relationship (R^2 = 0.925) of experimental activity with calculated activity for SVM aided linear tri-variable model. Non-linear relationship of experimental activity with calculated activity for SVM aided non-linear tri-variable model is presented in figure-2c. R^2 value (0.954) shows good predictive power of model.

Y-Randomization: In order to avoid by chance modelling we have performed Y-randomization method by repeated scrambling of biological activity. Each model was undertaken to 100 times replicate runs. Low values of correlation coefficient for all 100 models derived from Y-scrambling recommend that the generated model is not by chance. Graphical representations of Y-randomization for MLR tri-variable model and SVM aided linear and non-linear tri-variable models are reported in figure-3.

Descriptor sensitivity analysis: Most sensitive descriptor was evaluated by descriptor sensitivity analysis [25]**.** Descriptor GATS3m was found as most sensitive descriptor with high area under the curve (2.839) in multiple linear regression models. In SVM aided linear and non-linear QSAR models descriptor IC1 was found as the most sensitive descriptor with area under the curve respectively (7.810 and 0.249).

Conclusion

In the present study we have compared the performance of MLR and SVM aided linear and non-linear regression in QSAR models. Models obtained from SVM aided non-linear regression were found statistically fit and more predictive than models obtained from multiple linear regression and SVM aided linear regression. Descriptors (IC1, SIC3) and (GATS3m, GATS5p), used in MLR and SVM aided linear and non-linear regression are from same class of descriptors. These descriptors contribute to structure-activity relationship and code for same chemical structure property but differ in linear and non-linear relationship. Descriptor IC1 was found as the most sensitive descriptor with highest area under the curve. Descriptor sensitivity analysis illustrates the importance of molecular descriptors.

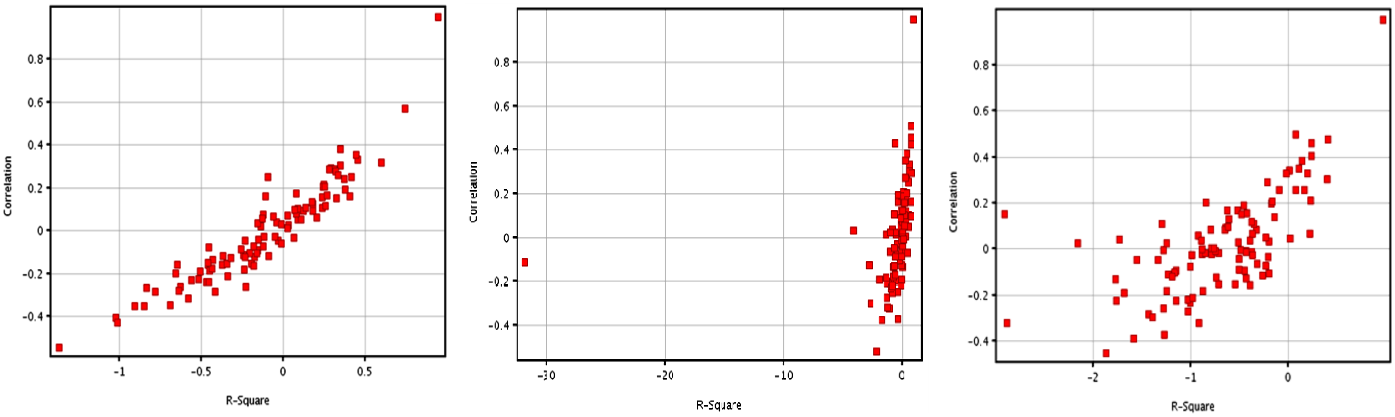

Figure-3
Y-Scrambling graphs for (a) MLR aided tri-variable model, (b) SVM aided linear tri-variable model, (c) SVM aided non-linear tri-variable model

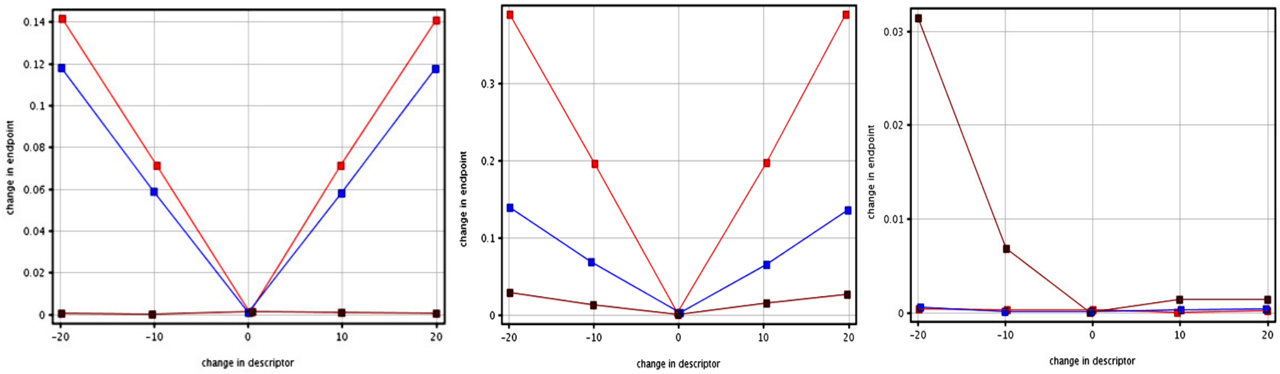

Figure-3
Descriptor sensitivity (a) MLR aided tri-variable model, (b) SVM aided linear tri-variable model, (c) SVM aided non-linear tri-variable model

References

1. Singh S., The Genetics of Type 2 diabetes mellitus: A Review., *J. Sci. Res.*, **55**, 35-48 **(2011)**

2. Qaseem A., Humphrey L.L., Oral Pharmacologic Treatment of Type 2 Diabetes Mellitus: A Clinical practice Guideline from the American college of Physicians, *Ann Intern Med.*, **156,** 218-231 **(2012)**

3. Bahare R.S., Gupta J., Malik S., Sharma N., New Emerging Targets for Type-2 Diabetes, *Intl. J. Pharm Tech.,* **3(2)**, 809-818 **(2011)**

4. Johnson T.O., Ermolieff J., Jirousek M.R., Protein tyrosine phosphatise 1b inhibitors for diabetes, *Nat. Rev. Drug Discov.*, **1,** 696-709 **(2012)**

5. Koren S, Fantus I.G., Inhibition of the protein tyrosine phosphatase PTP1B: potential therapy for obesity, insulin resistance and type 2 diabetes mellitus, *Best Pract Res Clin Endocrinol Metab*., **21**, 621-40 **(2007)**

6. Goldstein B.J., Protein-tyrosine phosphatase 1B (PTP1B): a novel therapeutic target for type 2 diabetes mellitus, obesity and related states of insulin resistance, Curr *Drug Targets Immune Endocr Metabol Disord*. **1,** 265-75 **(2001)**

7. Rosenbloom A.L., Silverstein J.H., Amemiya S., Zeitler P., Klingensmith G.J,. Type 2 diabetes in the child and adolescent, *Pediatr Diabetes*., **9**, 512–526 **(2008)**

8. Dearden J.C., In silico prediction of drug toxicity. *J. Comput Aided Mol. Des*.,**17**, 119 -27 **(2003)**

9. Nantasenamat C., Isarankura-Na-Ayudhya C., Naenna T., Prachayasittikul V., "A practical overview of quantitative structure-activity relationship". *Excli J.,* **8**, 74-88 **(2009)**

10. Nantasenamat C., Isarankura-Na-Ayudhya C., Naenna T., Prachayasittikul V., Advances in computational methods to predict the biological activity of compounds, *Expert Opin Drug Discov.,* **5**, 633–54 **(2010)**

11. Poole D., Mackworth A., Goebel R., Computational Intelligence: A Logical Approach, Oxford University Press, USA, **(1998)**

12. Yegnanarayana B., Artificial neural networks, PHI Learning Pvt. Ltd. (2009)

13. Joachims T., Text categorization with support vector machines: Learning with many relevant features. Technical Report 23, LS VIII, University at Dortmund **(1997)**

14. Rokach L., Maimon O., Data mining with decision trees: theory and applications. World Scientific Pub Co. Inc. **(2008)**

15. Ramoni M., Sebastiani P. "Robust bayes classifiers." Artificial Intelligence., **125**: 209-226 **(2001)**

16. Montgomery D.C., Peck E.A., Vining G.G., Introduction to linear regression analysis, John Wiley & Sons. 821 **(2012)**

17. Navarrete-Vazquez G. et al. Synthesis, in vitro and computational studies of protein tyrosine phosphatise 1B inhibition of a small library of 2-arylsulfonylamino benzothiazole with antihyperglycemic activity. *Bioorg Med chem.,* **17**:3332-3341 **(2009**

18. MarvinSketch version 5.5.1,**(2009)**, Chemaxon.http://www.chemaxon.com.

19. VCCLAB, Virtual Computational Chemistry Laboratory, http://www.vcclab.org. **(2005)**

20. Sarchitect^{TM} 2.5 Designer/Miner, Strand Life Sciences Pvt. Ltd., Bangalore, India, **(2008)**.

21. Cortes C., Vapnik V., Support-Vector Networks. Mach. Learn. **20**, 273-297 **(1995)**

22. Mangasarian O.L., Musicant D.R., Large scale kernel regression via linear programming, *Mach. Learn.,* **46**:255-269 **(2002)**

23. Klopman G., Kalos A.N., Causality in Structure-Activity Studies. *J. Comput. Chem*., **6**:492-506 **(1985)**

24. Wold S, Eriksson L. Statistical Validation of QSAR Results. In: van de Waterbeemd, H. (Ed.) Chemometric Methods in Molecular Design, Weinheim., 309-318 **(1995)**

25. Saltelli A., Tarantola S., Campolongo F., Ratto, M., Sensitivity analysis in practice: a guide to assessing scientific models. John Wiley & Sons, **(2004)**