Sensitivity Analysis for Environmental Models and Monitoring Networks

Alessandro Fassò

Dept. IGI, University of Bergamo, Italy - alessandro.fasso@unibg.it

Abstract: Statistical sensitivity analysis is shown to be a useful technique for assessing both multivariate environmental computer models and environmental statistical spatio-temporal models in the perspective of risk assessment. Methods are reviewed and extended to cover with two applications which are reported as case studies. The first, related to waste water biofilters for heavy metals, is aimed at assessing the input influence on both environmental and economical variables. The second, related to spatio-temporal models for air quality monitoring networks, is intended to study the influence of each station to the model performance.

Keywords: Heterogeneous Networks, space time models, calibration, particulate matters, waste water, heavy metals, biofilters

1 INTRODUCTION

Environmental models are often useful tools for environmental risk assessment. Consider, for example, a waste water biofilter for heavy metals. The assessment of the lifetime, rate of unfiltered heavy metals and safety are of great concern and having at disposal a computer model for simulation may relevantly improve risk assessment, environmental protection and plant management.

As a second example, consider environmental and human risks related to air quality pollution and the importance of monitoring and reconstructing the air quality field related to a heavily polluted region such as the Po Valley, Italy. This area is characterized by mountains in the south, west and north and a see coast in the east. The central plain is densely inhabited with heavy vehicle traffic, heating and industrial emissions. Moreover, intensive agricultural activity is an important source of land and water contamination.

In this paper, we consider two kinds of models, namely computer models (CM) and statistical models (SM). We say that a model is a CM if it has been deduced from some theory or calibrated and validated on enough empirical evidence and is now available in form of computer code. We are then interested in studying the influence of certain inputs to the model outputs and the propagation of the input uncertainty through the model. In this sense we are not directly interested in model uncertainty intended as the coherence of the model with reality which is taken for granted. In this paper, SA is essentially based on a certain variance decomposition and on the appropriate sampling plan allowing to estimate the related variance components, see e.g. Saltelli et al. [2004] and references therein.

On the other side, a SM is a simplified model which has to be calibrated on available data and try to describe the main features of a phenomenon under study. In this case, SA is interested in assessing the influence of the data on the model parameter and performance. Of course this distinction does not cover all applications and in some cases we may have CM with model uncertainty or CSM.

This paper reviews and extends results on both above approaches which have been recently obtained by the author and his collaborators and is organized as follows. Section 2 briefly reviews SA for CM's and gives an example on the previously mentioned waste water biofiltering. Section 5 discusses the extension of SA to spatio-temporal models for air quality monitoring networks and, considering the Po Valley monitoring network, shows how SA may help in both assessing the model performance and deciding on model order.

To make the paper self-contained, section3 gives concepts on SA for SM and, section 4, summarizes the main aspects of spatio-temporal modelling based on the GDC approach. A section of concluding remarks closes the paper.

2 SA OF COMPUTER MODELS

Uncertainty assessment of CM's may focus on various uncertainty sources as discussed in Kennedy and Hogan [2001] and Fassò [2006] . In this paper, we focus on that part of SA which is intended to rank model inputs for their influence to the CM's output. To see this consider a CM as a function $y = f(x)$ which relates the input parameter set $x = (x_1, ..., x_h)$ to the model output set y. In some cases it may be useful to use an *emulator* of the CM, which is a simplified statistical model of the CM. With some abuse of notation, if we denote also the emulator by $f()$ we get the new CM-emulator equation

$$y = f(x) + \varepsilon \tag{1}$$

where ε is the emulation error.

The main idea of SA is to assess the input influence using the output variance quota attributed to each input obtained by some variance decomposition. For scalar output y, a first order decomposition gives

$$Var(y) = \sum_j Var(y|x_j) + \sigma^2 \tag{2}$$

which can be computed using standard regression formulas and simple sampling plans if the conditional expectations $f_j = E(y|x_j)$ are approximately linear. If f_j are not linear, more complex plans are required but equation (2) still holds and a first order sensitivity index may be based on

$$S_j = \frac{Var(y|x_j)}{Var(y)}.$$

The simple decomposition of equation (2) may be extended in various directions. First, we may be interested in considering interactions of any order by a possibly orthogonal decomposition of f, namely

$$z = f_0 + \sum_j f_j + \sum_{i<j} f_{i,j} + ... + f_{j_1,...,j_k} \tag{3}$$

which yields an immediate extension of equation (2). Of course, complexity of representation (3) and the complexity of the corresponding sampling plan is related to computational complexity and computing time. So, on the one side, we can distinguish between cheap or not time consuming code and expensive or computer intensive code and, on the other side, we have preliminary analysis design of experiments, Monte Carlo sampling, improved Monte Carlo plans and Bayesian designs.

A second direction considers multivariable CM's with $y = (y_1, ..., y_k)$. In this case $V_y = Var(y)$ is a variance-covariance matrix and its trace or its determinant have been used in literature. Here, we propose a sensitivity index based on the variance decomposition of the linear combination $\alpha' y$. To see this, suppose that the $emulator$ is in the form $y = Bx + \varepsilon = \Sigma_j B_j x_j + \varepsilon$ where B is a $k \times h$ matrix with j^{th} column B_j and

$$\alpha' V_y \alpha = \alpha' B V_x B' \alpha + \alpha' V_\varepsilon \alpha.$$

Hence, if x has uncorrelated components with $V_x = diag\left(\sigma^2_{x_1}, ...\sigma^2_{x_k}\right)$, the *quadratic* sensitivity index

$$S^\alpha_j = \sigma^2_{x_j} \frac{\alpha' B_j B'_j \alpha}{\alpha' V_y \alpha} \tag{4}$$

takes into account the correlation among the components of y and retains additivity as $R^2_{\alpha' y} = 1 - \frac{\alpha' V_\varepsilon \alpha}{\alpha' V_y \alpha} = \Sigma_j S^\alpha_j$.

The third direction for extending decomposition (2) is heteroskedasticity. According to this, the emulation error from equation (1) has a variance which depends on x.

2.1 SA of a Biofilter for Heavy Metals

In this example, we re-consider the SA of a waste water biofilter for heavy metals by means of a CM for biosorption in packed column reactors discussed in Fassò et al. [2003].

The computer model $y = f(x)$ given by the numerical solution of a PDE, has two outputs for assessing the performance of the fixed bed column, namely the breakthrough time (t_b) and the length of unused bed (LUB). The former, being the column working time over which the outlet metal concentration exceeds 5% of the inlet concentration, is related to environmental risk of polluted water discharge. The latter, representing the column efficiency, is related to economical aspects. Both are of concern for plant design and management.

The input set x is 8-dimensional with the first three components which are the fluid dynamic factors: liquid viscosity (μ_L), liquid density (ρ_L), specific bed velocity (u_0). Whereas the remaining five components are chemical-physical characteristics, namely: column void degree (ε), adsorption particle diameter (d_p), density of the biosorbent (ρ_S) and adsorption characteristics given by the maximum intake q_{max} and the Langmuir constant b.

The mentioned SA of Fassò et al. [2003] was a *global* investigation, exploring a large range of the input space and covering for nonlinearities and heteroskedasticity. In this section, we consider a Monte Carlo *local* analysis using a reduced subset of the parameter space with ranges of each input of about 25% of the previous ones. Whereas the global investigation where aimed at preliminary analysis and plant design, the local SA may be useful for final plant implementation and management. The Monte Carlo sample is summarized in Table 1 and is composed of $n = 1'000$ replications from the independent uniform distribution except for $q_{\max}$ and b, which come from the bivariate Gaussian distribution, and u_0, which is kept constant.

	Min	Max	Mean	Std
μ_L	0.55	0.65	0.60	0.03
ρ_L	0.95	1.05	1.00	0.03
ε	0.25	0.35	0.30	0.03
d_p	0.07	0.13	0.10	0.01
ρ_S	1.05	1.15	1.10	0.03
$q_{\max}$	29.36	52.87	40.07	3.97
b	0.46	4.44	2.50	0.63
t_b	0.37	0.53	0.45	0.03
LUB	0.10	0.22	0.16	0.02

Table 1: Statistics for Monte Carlo sample $n = 1'000$

Due to the local nature of this study it is not surprising that a bivariate linear emulator is quite satisfactory with Gaussian errors and very high fitting $R^2 = 0.99$ for both t_b and LUB components. Table 2 gives the sensitivity indexes for both the two univariate models and the multivariate one. Note that the maximum uptake $q_{\max}$ is the most important input in this range for the breakthrough time t_b and hence for environmental protection filter regeneration policy. Since t_b and LUB are negatively correlated with correlation coefficient $r = -0.39$, the old SI based on the trace of the covariance matrix and the new S_j^α of equation (4) with $\alpha' = (1, 1)$ give quite different results.

	Univariate SI		Multivariate SI	
	t_b	LUB	trace	S_j^α
$q_{\max}$	88.0	29.6	58.8	12.8
ε	4.9	61.7	33.3	83.7
ρ_S	6.1	2.1	4.1	0.9
b	0.8	5.8	3.3	1.9
d_p	0.0	0.2	0.1	0.1

Table 2: Percentage Sensitivity Indexes

As a by-product of this SA, we have the uncertainty empirical distribution of t_b and LUB which can be used for computing the probability of a large discharge before the regeneration time.

3 SA OF STATISTICAL MODELS

In fitting statistical models to data, it is often of interest to assess the influence of each observation on the model performance. Often a *statistical* (parametric) model for y has the following structure

$$y = m(x, \theta) + h(x, \theta)\,\varepsilon$$

where ε is a Gaussian white noise, m is some forecasting function and h a skedastic function, x is a set of known regressors and θ is an unknown model parameter to be estimated.

The model is statistical in the sense that, given some data $Y = (y_1, ..., y_n)$, we get an estimate of θ, namely $\hat{\theta} = \hat{\theta}(y_1, ..., y_n)$. The influence of an observation, say y_i, on the model may be assessed by considering the estimate $\hat{\theta}$ when we omit y_i and use $Y_{-i} = (y_j, j = 1, ..., n, j \neq i)$ namely

$$\hat{\theta} - \hat{\theta}_{-i} = \hat{\theta}(Y) - \hat{\theta}(Y_{-i}). \tag{5}$$

In some other cases, we are interested in assessing the influence of y_i to the *forecasting performance* of the model, for example using

$$\hat{m} - \hat{m}_{-i} = m\left(x, \hat{\theta}\right) - m\left(x, \hat{\theta}_{-i}\right). \tag{6}$$

Analytical study of quantities derived by (5) or (6) can be done in simple cases and the influence on $\hat{\theta}$ may be done by the so-called *influence curves,* see e.g. Fassò and Perri [2002] .

If m is a standard multiple linear regression model with constant skedastic function $y = \theta' x + \varepsilon$ and independent observations $y_1, ..., y_n$, then it is known that the Cook's distance is connected with both (5) and (6) , namely

$$\begin{aligned} D_{-i} &= c\left(\hat{\theta} - \hat{\theta}_{-i}\right)' \hat{\Sigma}^{-1}\left(\hat{\theta} - \hat{\theta}_{-i}\right) & (7) \\ &= c' \sum \left(\hat{m}(x_j) - \hat{m}_{-i}(x_j)\right)^2 & (8) \end{aligned}$$

where c and c' are constants and $\hat{\Sigma}$ is the estimated covariance matrix of $\hat{\theta}$. Note that D_{-i} has an unknown distribution, nevertheless it is common practice to use some χ^2 percentiles as reference values.

For more complex models such as time series models or spatio-temporal models, expressions (7) and (8) are not equivalent and the first one is used for model sensitivity whilst, for forecasting performance, it can be used the mean absolute error $MAE_{-i} = \frac{1}{n}\sum_j |y_j - \hat{m}_{-i}(x_j)|$ or the corresponding root mean squared error $RMSE_{-i}$, and model bias $b_{-\imath}$.

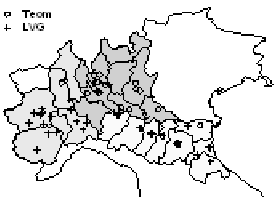

Figure 1: The Po Valley PM_{10} monitoring network.

4 MONITORING MODELS

In this section, in order to give a frame for the sensitivity analysis of the next section, we briefly discuss the application of GDC model of Fassò et al. [2004] , [2005] and [2006] to the one year air quality monitoring data on PM_{10} daily concentrations coming from the Po Valley network, Italy, which consists of the $n = 54$ sites depicted in Figure 1. Since, we have a marked instrumental heterogeneity with so-called $TEOM$ and LVG monitors, measurement equations for observations from these two instruments, namely y_G and y_T, at locations $s = s_1, ..., s_n$ and time $t = 1, 2, ..., N = 365$, are given by

$$\begin{array}{l} y_G(t,s) = y^*(t,s) + \varepsilon_G(t,s) \\ y_T(t,s') = \alpha + \beta y^*(t,s') + \varepsilon_T(t,s'). \end{array} \tag{9}$$

Here, $y^*(t,s)$ are realizations of the discrete-time continuos-space process, which can be considered as the underlying *true* pollution level at time t and location s. Moreover, $y^*(t,s)$ is supposed to be a linear function of a common *regional* k-dimensional process denoted by $\mu(t)$, namely

$$y^*(t,s) = \Phi(s)\,\mu(t) \tag{10}$$

where the loadings matrix $\Phi = (\Phi(s_1)', \ldots, \Phi(s_n)')'$ is obtained by a k-dimensional set of Empirical Orthogonal Functions (EOF) or PCA decomposition of the covariance matrix of $(y(t,s_1), ..., y(t,s_n))\,, t = 1, ..., N.$

The underlying regional process $\mu(t)$, which has a dimensionality appropriated to cover with the spatial complexity, has Markovian dynamics given by

$$\mu(t) = H\mu(t-1) + \eta(t)\,. \tag{11}$$

The innovation process η is a $k-$dimensional zero mean Gaussian white noise. Its covariance matrix Σ_η and Markovian propagation matrix H are both diagonal with $\Sigma_\eta = diag(\sigma^2_{\eta_1}, ..., \sigma^2_{\eta_k})$ a s a natural consequence of the EOF approach and using the homogenous propagation hypothesis $H = diag(h, ..., h)$. Moreover the error component $\varepsilon(t,s)$ is a Gaussian spatially and time independent process, with sphericity assumption, $\Sigma_{\varepsilon(t)} = diag(\sigma^2_\varepsilon, ..., \sigma^2_\varepsilon)$.

4.1 Parameter Estimation

Conditionally on the loading matrix Φ of the previous subsection with $k = 14$ principal functions, we estimate the GDC model parameter set θ using the maximum likelihood method as in Table 3.

The model parameter vector θ includes the calibration constants from measurement equation, namely α, and β, the propagation coefficient h. Moreover, we have innovation variances, namely $\Sigma_\eta = diag(\sigma^2_{\eta_1}, ..., \sigma^2_{\eta_k})$ with $\sigma^2_{\eta_5} = ... = \sigma^2_{\eta_k}$ and, finally, the measurement error variance σ^2_ε

α	β	h	θ_{η_1}	θ_{η_2}	θ_{η_3}	θ_{η_4}	θ_ε
2.32	0.34	0.996	2.16	-0.38	-0.52	-4.34	-1.28
0.03	0.007	0.002	0.08	0.15	0.16	0.12	0.01

Table 3: MLE's and their standard deviations for GDC model. Notation: $\theta_{\eta_j} = \log\left(\sigma^2_{\eta_j}\right)$

5 SA OF MONITORING MODELS

In order to evaluate the sensitivity to the network configuration, we carry out a cross validation analysis by removing one station at a time. In particular, after excluding the i-th location from the data input, we compute the loading matrix, Φ_{-i} say, and using the GDC model $(9) - (11)$, we estimate the corresponding parameter vector, θ_{-i} say.

5.1 Parameter Sensitivity

Figures 2 and 3, show the influence of each station to the instrument additive bias α_{-i} and the calibration coefficient β_{-i}. These and the remaining figures are divided into two areas, the left one is for LVG monitors and the right one for $TEOM$ monitors. Moreover, we add a solid line referring to the parameter value of the general GDC model of Table 3 and dashed lines for the $\pm 2\sigma$ interval, where the estimated standard deviation of the MLE is used for σ. It is apparent that, some stations are very influential to these parameters. Moreover α_{-i} and β_{-i} are strongly negatively correlated as the MLE and the cross validation procedure give a correlation coef-

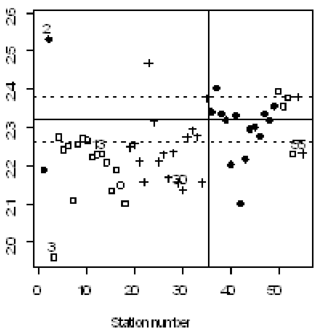

Figure 2: Instrument additive bias α_{-i} for LVG (left) and TEOM (right). Legend: solid circle: Lombardia; square: Emilia Romagna; cross: Piemonte. Solid line: coefficient α estimated from the full network (see Table 3); dashed lines: $\alpha \pm 2\sigma_{\alpha}$.

ficient amounting to -0.93 and -0.85 respectively. Hence their joint analysis should adjust for this.

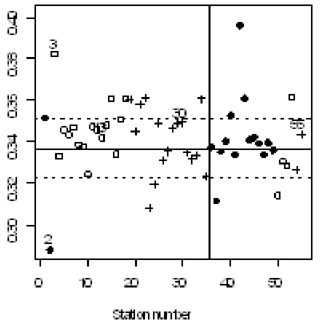

Figure 3: Instrument multiplicative bias β_{-i}. Legend: see Figure 2

On the other side Figure 4 shows that, the network has a negligible influence on the Markovian persistence parameter h. To take into account all the parameters together and their mutual correlation, we assess the influence of each station to the model as a whole by using the *Cook distance* (7) which extends sensitivity analysis, of section 3 to heterogeneous networks.

5.2 Prediction Sensitivity

Since the predictions of y_G and y_T at site s_i are given by $\Phi(s_i)\mu(t)$ and $\alpha + \beta\Phi(s_i)\mu(t)$ respectively, we calculate the daily cross validation errors for each station, as differences between the predicted and the observed data. Subsequently, the station bias given by the yearly average error for each station is drawn in Figure 7; this and Figure 6, which reports the MAE, allow us to assess the reconstruction capability of the model. Note that, as often happens, largest values of MAE are associated with the largest Bias.

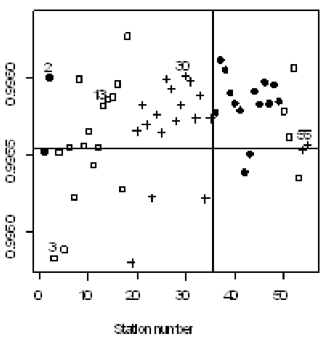

Figure 4: Propagation coefficient h_{-i}. Legend: see Figure 2.

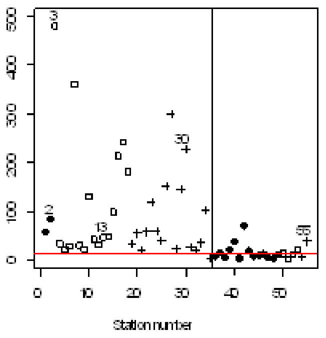

Figure 5: Cook distance D_{-i}. Legend: see Figure 2. Solid line: 95° percentile from χ^2 distribution with $\dim(\Omega)$ degrees of freedom.

Moreover, the comparison of these figures with Cook distance and related Figures 2, 3, 4 and 5 allows us to discriminate between stations which have influence on parameter estimation and spatial prediction. For example, Station n.3, located in Piacenza and called the *Pubblico Passeggio Station*, has extreme values in all the considered diagnostic graphical tools. Hence it is quite influential to both parameter estimation and data reconstruction and, in this sense, it could be considered as an outlier. Average $bias$, MAE and $RMSE$ for various PCA dimensions k are reported in Table 4. Being network overall values, these quantities give the

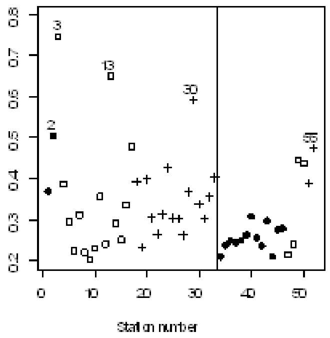

Figure 6: Mean Absolute Error MAE_{-i}. Legend: see Figure 2.

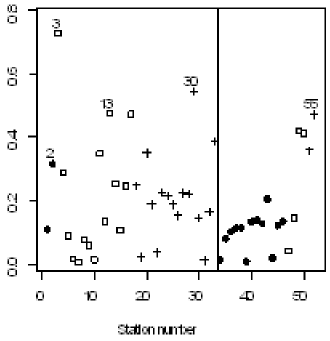

Figure 7: Network Bias b_{-i}. Legend: see Figure 2

sensitivity of the model performance to the parameter k and justify the value $k = 14$ used in section 4.1.

	k				
	6	10	14	18	22
RMSE	0.446	0.451	0.434	0.445	0.433
MAE	0.345	0.348	0.330	0.339	0.330
Bias	0.206	0.220	0.197	0.199	0.193

Table 4: Network crossvalidation performance for PCA dimension k.

6 CONCLUSIONS

Sensitivity Analysis of multivariate computer code requires special attention because output correlation may amplify or reduce single component sensitivity indexes. The quadratic index proposed in this paper seems to be able to cope with such problem.

Spatio-temporal models for air quality may be heavily influenced by the network configuration. The extension of Cook's distance approach to such models enables as to assess such influence and improve network design and understanding.

REFERENCES

Fassò A., Sensitivity Analysis and Water Quality. *Working Paper GRASPA*, **23**, (www.graspa.org). In printing on: Wymer L. Ed, *Recreational Beaches: Statistical Framework for Water Quality Criteria and Monitoring*. Wiley. 2006.

Fassò A., Cameletti M. and Nicolis O. Air quality monitoring using heterogeneous networks, submitted, 2006.

Fassò A., Esposito E., Porcu E., Reverberi A.P., Vegliò F. Statistical Sensitivity Analysis of Packed Column Reactors for Contaminated Wastewater. Environmetrics. **14**(8), 743 - 759, 2003.

Fassò A, Nicolis O. Modelling dynamics and uncertainty in assessment of quality standards for fine particulate matters. *Working Paper GRASPA,* **21**, (www.graspa.org), 2004.

Fassò A, Nicolis O. Space-time integration of heterogeneous networks in air quality monitoring. *Proceedings of the Italian Statistical Society Conference* on "Statistica e Ambiente", Messina, 21-23 September 2005", 1, 2005.

Fassò A., Perri P.F. Sensitivity Analysis. In Abdel H. El-Shaarawi and Walter W. Piegorsch (eds) Encyclopedia of Environmetrics, **4**, 1968–1982, Wiley, 2002.

Kennedy M.C. O'Hagan A. Bayesian calibration of computer models, *J. Royal Stat. Soc. B*, **63**, 425-464. 2001.

Saltelli A., Tarantola S., Campolongo F., Ratto M. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, Wiley. 2004.