Nonlinear Multivariate Statistical Sensitivity Analysis of Environmental Models

with Application on Heavy Metals Adsorption from Contaminated Wastewater

A. **Fassò**, **A**. **Esposito**, **E**. **Porcu**, **A**.**P**. **Reverberi**, **F**. **Vegliò**
Working paper & contact: http://**ingegneria**.**unibg**.**it/fasso**

Work partially supported by italian MURST grant.

Summary

In this talk we consider a computable function

$$y = f(\mathbf{x})$$

or computer model which describes some environmental system.

Sensitivity Analysis (SA) aims to assess the **uncertainty** on y and the **importance** of each **input** or **parameter** x_j from

$$\mathbf{x} = (x_1, .., x_k)$$

We extend standard results from scalar y to vector valued y and extend from linear $\hat{f}$ to heteroskedastic models.

Contaminated wastewater treatment application

We apply these methods to SA of fixed bed reactor in order to assess the influence of packed column parameters on the heavy metal biosorption performance.

Talk Plan

1. Statistical Method
- *SA* Introduction
- Extension 1: Multivariate *SA*
- Extension 2: Heteroskedastic *SA*

2. Application
- Packed Column Computer Model
- MC Simulation of Biosorption
- Data Modelling & *SA*
- Conclusions

Statistical SA

SA Methods are extensively considered in Saltelli et al. (2000) and reviewed by Fassò and Perri (2001).

In defining importance measures we have two main streams in SA :

1. Non-parametric: getting conclusions without much emphasis on $f(\quad)$
2. Parametric: try to get a simple but accurate and phisically sound statistical model for $f(\quad)$

Following the latter we use and extend **regression based Monte Carlo** SA

Basic idea:

- Use **probability distributions** for modelling uncertainty on
  - input
  - output
- The joint pdf of **x**, $p(\mathbf{x})$ say, is known
- The output pdf is to be estimated
- Importance measures are based on some Variance decomposition technique.
- Get a sample from $p(\mathbf{x})$, i.e. n repeated stochastic simulations of **x**, and n computer runs, giving:

$$\mathbf{z}_i = (y_i, \mathbf{x}_i), \qquad i = 1, ..., n$$

These data can be used in a statistical framework to get a simplified version $\hat{f}$ such that

$$y = \hat{f}(\mathbf{x}) + e$$

- the statistical model $\hat{f}$ can be used for insight into the system dynamics and in particular for assessing the importance of the various parameters.
- In order to use a linear regression, we use a post-simulation input transform

  $$\mathbf{u} = u(\mathbf{x})$$

  which may give:
  - zero mean incorrelated inputs
  - polynomial, interactions and other nonlinearities

  We get the transformed problem model:

  $$\tilde{y} = \boldsymbol{\beta}' \mathbf{u} + e$$

  with simulated zero mean deviates $\tilde{y} = y - \bar{y}$ (The tilde will be omitted for simplicity).

- When the components u_j are incorrelated, the output variance σ_y^2 can be decomposed as

$$\sigma_y^2 = \sum_j \beta_j^2 \sigma_{u_j}^2 + \sigma_e^2$$

  and using the LS estimate $\hat{\beta}_j$ a similar decomposition holds for sample variances $S_y^2, S_{u_j}^2$ and S_e^2 (computed with the same denominator e.g. $\frac{1}{n-1}$).

- It follows that the a natural **importance measure** or **sensitivity index** is given by

$$SI_j = \hat{\beta}_j^2 \frac{S_{u_j}^2}{S_y^2}$$

  can be used to assess the **influence** of the j^{th} input to the model output.

- In fact these indexes sum to the total variance percentage explained by the model, i.e.

$$\mathbf{R}^2 = \sum_j SI_j = 1 - \frac{S_e^2}{S_y^2}$$

where $\mathbf{R}^2$ is the well known multiple determination coefficient and the parameters can be ranked accordingly.

Multivariate *SA*

For the sake of simplicity consider only the bivariate case, which is the case study of next section.

We have the following matrix notation for the multivariate regression model

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1' \\ \boldsymbol{\beta}_2' \end{pmatrix} \mathbf{u} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \mathbf{Bu} + \mathbf{e}$$

and the multivariate LS method (see e.g. Rencher, 1995) simply gives

$$\hat{\mathbf{B}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}(y_1) \\ \hat{\boldsymbol{\beta}}(y_2) \end{pmatrix}$$

where $\hat{\boldsymbol{\beta}}(y_1)$ is the ordinary LS estimate as in the previous section.

SA can now be applied to the variance-covariance matrix decomposition extending the scalar variance decomposition:

$$\mathbf{S}_y = \mathbf{B}\mathbf{S}_u\mathbf{B}' + \mathbf{S}_e.$$

In order to get scalar valued indexes some metric for matrices is required:

- trace
- determinant
- likelihood decomposition

Here we use the trace metric which retains additivity.

We then have

$$\begin{aligned} tr(S_y) &= S_{y_1}^2 + S_{y2}^2 \\ &= \sum_j SI_j(y_1) S_{y_1}^2 + \sum_j SI_j(y_2) S_{y_2}^2 + S_{e_1}^2 + S_{e_2}^2 \end{aligned}$$

natural multivariate sensitivity indexes for j^{th} input are the simple or the weighted averages:

$$SI_j(y) = \frac{SI_j(y_1) + SI_j(y_2)}{2}$$

$$SI_j(y) = \frac{SI_j(y_1)S_{y_1}^2 + SI_j(y_2)S_{y_2}^2}{S_{y_1}^2 + S_{y_2}^2}$$

The first formula is appropriate for scale invariant SA and is the same as the second one if we use standardized outputs with unit variance.

Heteroskedastic *SA*

In technological scalar regression models it may happen that the error variance is not constant but changes with some factors v_j.

In this case the model is called heteroskedastic and the error

$$e = y - f(\mathbf{x})$$

is now given by

$$e = \zeta h$$

$$\zeta = NID(0,1)$$

$$h^2 = \alpha_0 + \sum_j \alpha_j v_j$$

Coefficients β and α can be jointly estimated via Gaussian maximum likelihood or by iterated weighted LS.

In this case the ANOVA decomposition, similarly to mixed models, is extended to cope with random effect components, namely

$$\sigma_y^2 = \sum_j \beta_j^2 \sigma_{u_j}^2 + \sum_j \alpha_j E(v_j) + \alpha_0$$

when u_j and v_j refer to the same input the corresponding heteroskedastic sensitivity index, say HSI, is

$$HSI_j = SI_j + SI_j^*$$

with

$$SI_j^* = \hat{\alpha}_j \frac{\bar{v}_j}{S_y^2}$$

For example, consider the case

- $\mathbf{u} = \mathbf{x} - \bar{\mathbf{x}}$
- $v_1 = x_1.$
- and skedastic component given by

$$h^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2$$

It follows that the sensitivity index for the first input is given by

$$HSI_1 = SI_1 + SI_1^* = \frac{\beta_j^2 \sigma_{u_j}^2}{\sigma_y^2} + \frac{\alpha_1 E(x_1) + \alpha_2 E(x_1^2)}{\sigma_y^2}$$

whilst the other sensitivities are unchanged,

$$HSI_j = SI_j$$

for $j = 2, ..., k.$

SA of

Heavy Metals Adsorption from Contaminated Wastewater

Computer Simulation Results

We than have the following I/O definition

Model Output

$$y = (t_b, LUB)$$

with

- $t_b = \frac{t_b^* u_0}{L}$ and t_b^* is the column working time,
- $LUB = \frac{LUB^*}{L}$ and LUB^* is the length of unused bed.

Input Parameters

$$\mathbf{x} = (\mu_L, \rho_L, u_0, \varepsilon, d_p, \rho_S, q_{\max}, b)'$$

Fluid Dynamic Factors

- μ_L fluid viscosity ($kg/m \times s$),
- ρ_L liquid density (kg/m^3),
- u_0 specific bed velocity (m/s),
- ϵ void degree,

Chemical-physical characteristics

- d_p adsorption particle diameter (mm),
- ρ_s sorbent density (kg/m^3),
- $q_{\max}$ maximum heavy metal up-take (mg/g) (Langmuir equation),
- b equilibrium solute Langmuir equation coefficient (L/mg).

Monte Carlo Simulations

$n = 10,000$ replications of **x** were simulated using independent rectangular random numbers and model outputs $y = (LUB, t_b)$ —— were computed

	Min	Max	Mean	Std
μ_L	0.298	0.903	0.603	0.173
ρ_L	0.801	1.200	1.001	0.115
u_0	196.801	603.791	402.047	115.470
ε	0.199	0.402	0.300	0.058
d_p	0.045	1.004	0.521	0.274
ρ_S	0.694	1.602	1.153	0.260
q_{max}	19.812	60.079	39.808	11.547
b	0.972	4.028	2.505	0.866
t_b	0.228	0.743	0.450	0.102
LUB	0.059	0.422	0.186	0.060

Table 1: Parameter and output statistics for global SA

Data Analysis and Modelling

This large dataset used allows nonlinearities and interactions to be detected with high power.

The modelling philosophy has been incremental:

- starting from a simple linear model as with $\mathbf{u} = \mathbf{x}$,
- deleting unimportant variables and then
- adding quadratic and interactions when needed
- first univariate modelling than multivariate extension

Diagnostic tools

- graphical analysis of residuals.
- searching for residuals almost independent from the x and
- normally distributed or, at least symmetric around zero and of course with
- small Mean Squared Error (MSE) and
- little model complexity as measured by the Akaike Information Criterion (AIC).

Data details

- All variables, both input and output, in the sequel are zero mean deviates.
- Linear inputs **x** are incorrelated
- the quadratic and interaction components have been orthogonalized after running the model so that they have to be interpreted as quadratic effects after discounting for the linear ones.

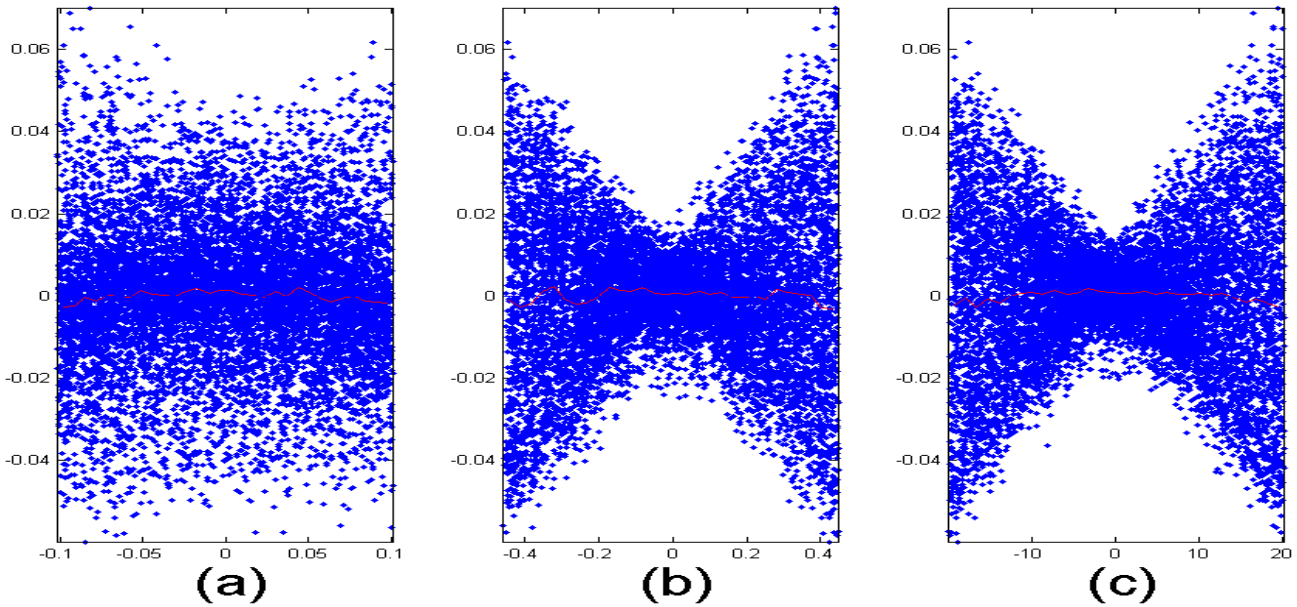
t_b Model 1

Omitting the statistically unimportant variables we get the following simple model for t_b, with $\pm standard\ deviation$ reported in brackets for all coefficients and error:

$$t_b = 0.187(\pm 0.003)\varepsilon + 0.237(\pm 7 \times 10^{-4})\rho_S + 0.00686(\pm 2 \times 10^{-5})q_{\max} + e(\pm 0.018)$$

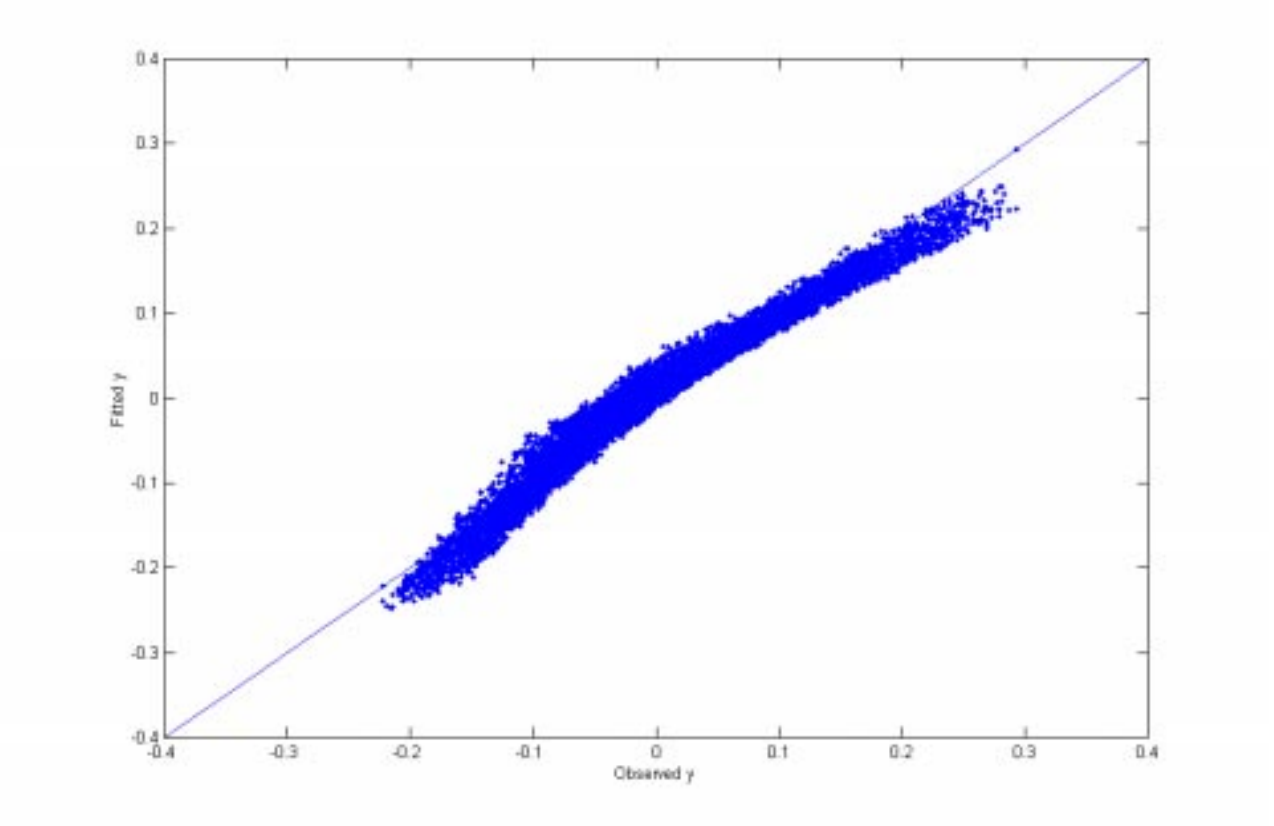
Model diagnostic

Very good fitting $R^2 = 96.90\%$, but ....



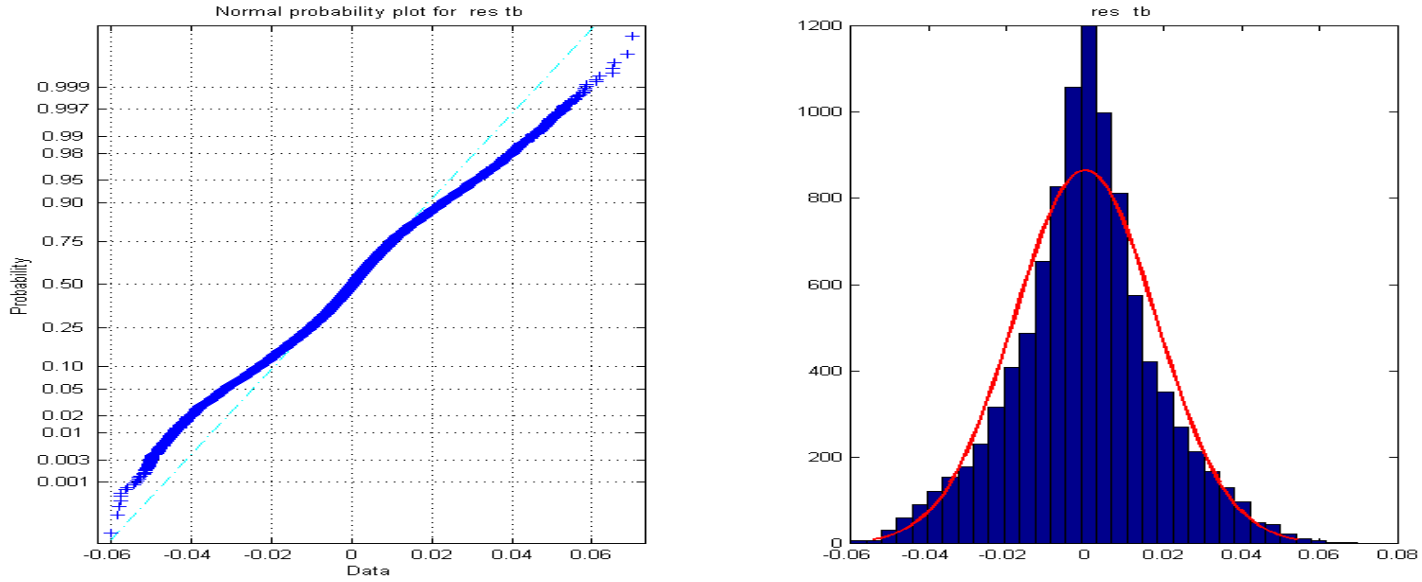
Residuals of t_b model 1 vs. parameters. (a) ε, (b) ρ_S, (c) $q_{\max}$.

The symmetric behavior of the residuals hints both for heteroskedasticity or interactions.



t_b values fitted by model 1 vs. observed.

Moreover the fitted against observed plot hints for nonlinearity



Residuals of t_b model 1. Normal probability plot and histogram

The same conclusions may be achieved from the normal probability plot and the histogram with superimposed Gaussian which show high tails and kurtosis

$$k(e) = \frac{n\sum e^4}{\left(\sum e^2\right)^2} = 3.5$$

t_b Model 2

We then get the following second order model

$$t_b = 0.187(\pm 9 \times 10^{-4})\varepsilon + 0.237(\pm 2 \times 10^{-4})\rho_S + 0.00686(\pm 10^{-5})q_{\max}$$
$$- 0.230(\pm 0.0033)\varepsilon\rho_S - 0.0071(\pm 2 \times 10^{-5})\varepsilon q_{\max} + 0.0054(\pm 2 \times 10^{-5})\rho_S q_{\max}$$
$$- 0.329(\pm 0.016)\varepsilon^2 - 0.00001(\pm 8 \times 10^{-6})q_{\max}^2 + e(\pm 0.0049).$$

Note that,

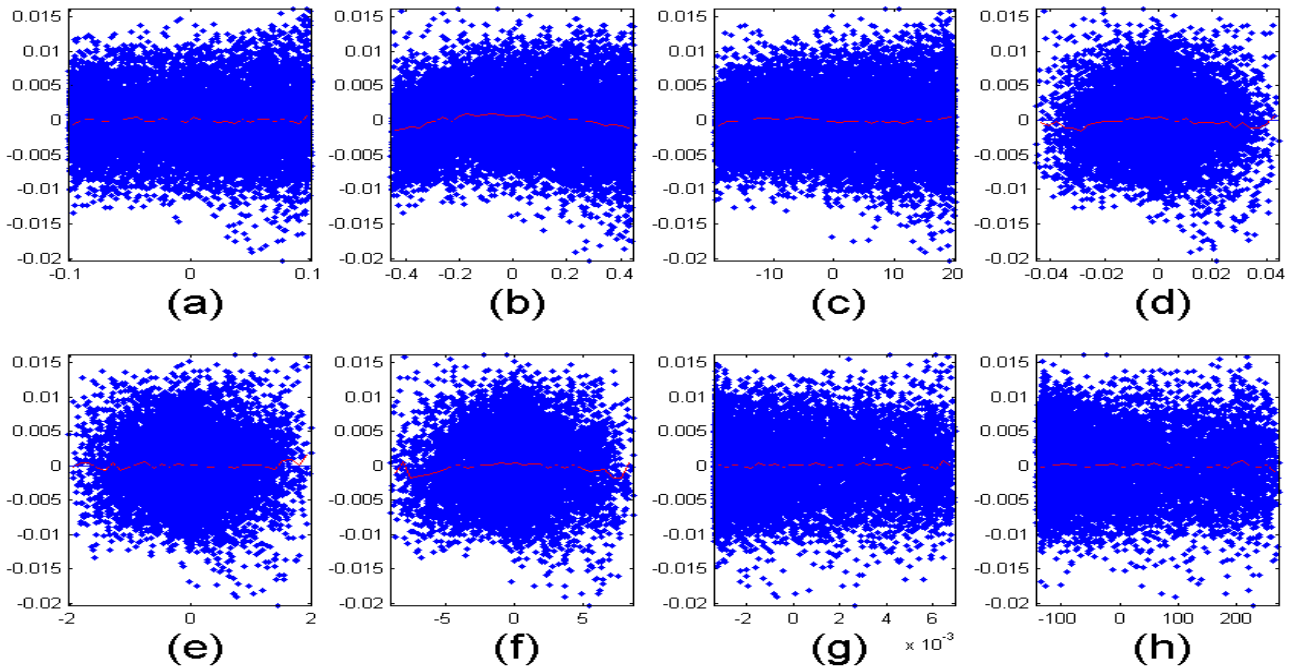
- input orthogonality $\Rightarrow$ the coefficients of the linear component are unchanged but all standard errors decreased
- fitting is now very high with $R^2 = 99.77\%$
- square root MSE is given by $RMSE = 0.015$.

Technical Details:

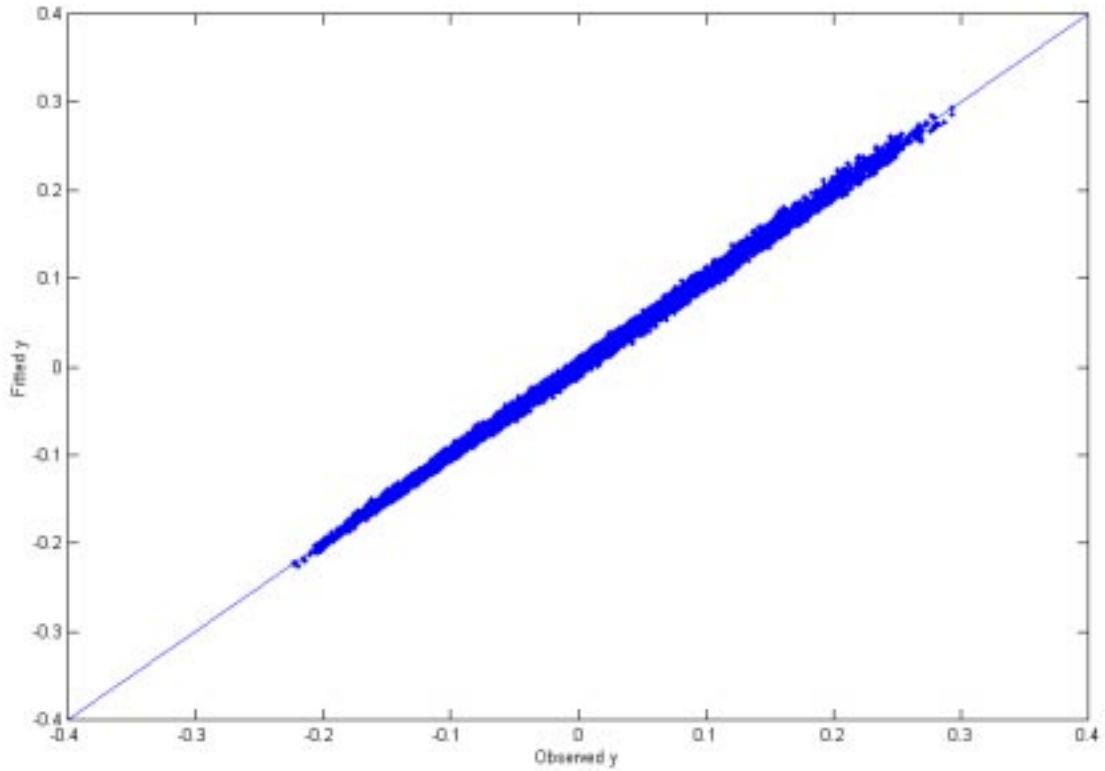
Remembering the non-stochastic nature of our simulation model one could search for a more detailed description of *f*( ).

For example, if one would use all 8 parameters and their 36 quadratic and interaction terms, would get $R^2 = 99.985\%$ and $RMSE = 0.0013$.

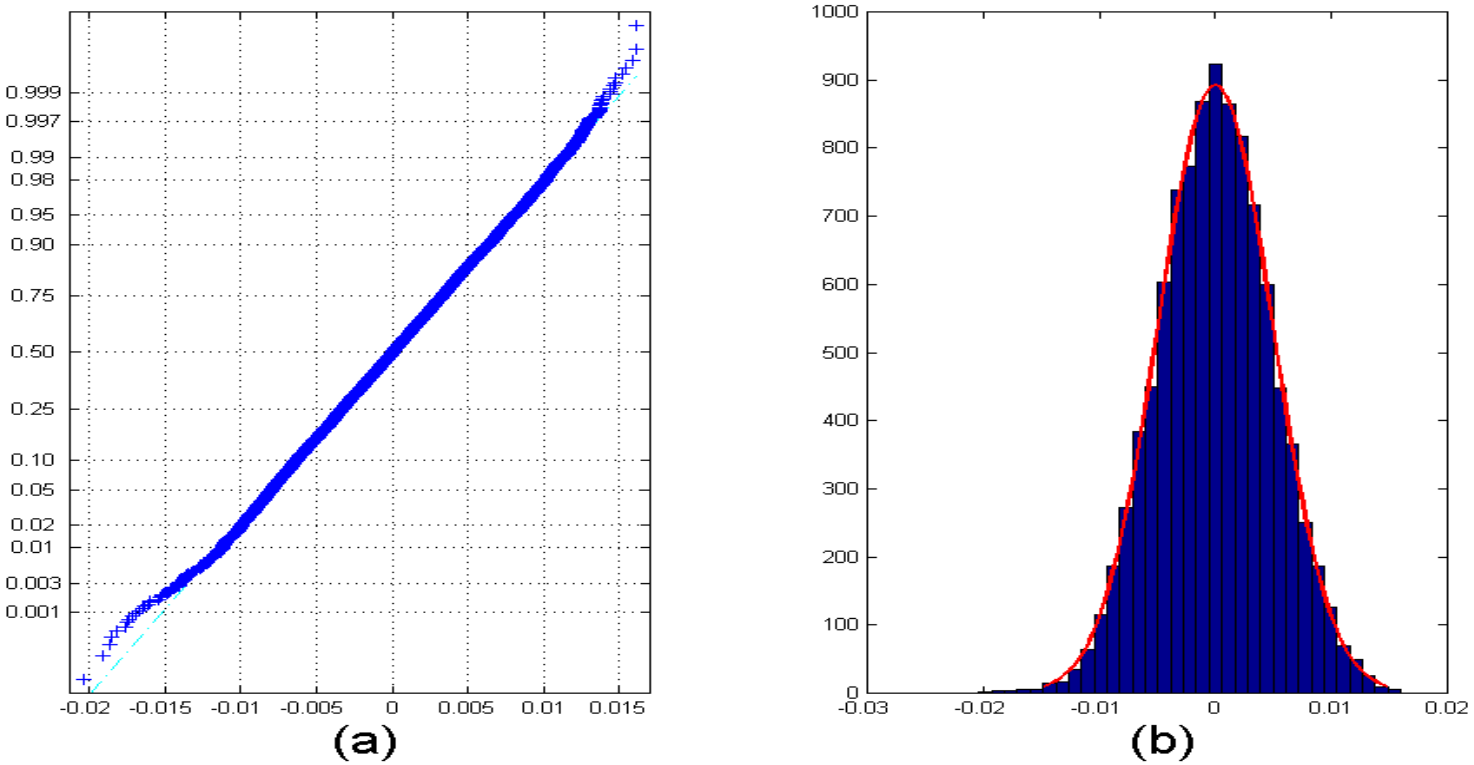
In this case, residual uncertainty would be about a quarter of t_b Model 2 but we omit these components because inessential for the present work.



Residuals of t_b model 2 vs. linear, quadratic and interaction terms. (a) ε, (b) ρ_S, (c) $q_{\max}$, (d) $\varepsilon\rho_S$, (e) $\varepsilon q_{\max}$, (f) $\rho_S q_{\max}$, (g) ε^2, (h) $q_{\max}^2$.



t_b values fitted by model 2 vs. observed.



Residuals of t_b model 2 : (a) normal probability plot; (b) .histogram

From these figures we see that we have an

almost perfect Gaussian error distribution

with

- no dynamical signal
- low skewness, $s_k = -0.03$
- gaussian kurtosis $k(e) = 3.03$.

Note that the pure quadratic components

$$\varepsilon^2 \text{ and } q^2_{\max}$$

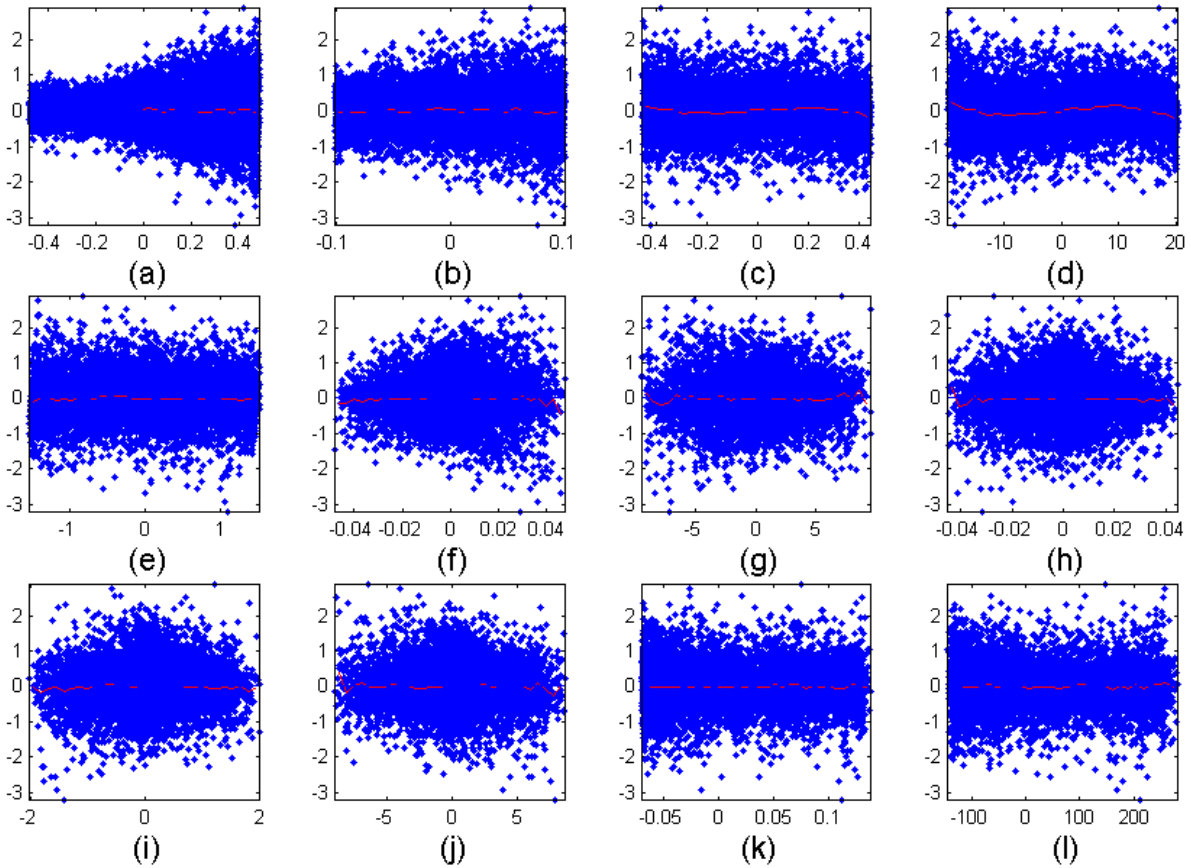
add very little in terms of variance but they have been retained in the model because improve both Gaussian fitting and dynamics filtering as shown by figures above

Modelling LUB

Following the approach of previous section we get the following second order model for $L = 100 \times LUB$

$$
\begin{aligned}
L = {} & 5.89(\pm 0.02) d_p + 62.6(\pm 0.098)\varepsilon - 10.0(\pm 0.02)\rho_S - 0.297(\pm 0.0005) q_{\max} \\
& + 0.722(\pm 0.006) b \\
& + 21.0(\pm 0.4) d_p \varepsilon - 0.0298(\pm 0.002) d_p q_{\max} \\
& - 18.0(\pm 0.4)\varepsilon\rho_S - 0.514(\pm 0.008)\varepsilon q_{\max} + 0.0776(\pm 0.002)\rho_S q_{\max} \\
& - 6.12(\pm 0.09)\rho_S^2 - 0.00527(\pm 5 \times 10^{-5}) q_{\max}^2 + e(\pm 0.56)
\end{aligned}
$$

This model again has a very good fitting with $R^2 = 99.12\%$.



Residuals of LUB model vs. linear, quadratic and interaction terms. (a) d_p, (b) ε, (c) ρ_S, (d) $q_{\max}$, (e) b, (f) $d_p\varepsilon$, (g) $d_p q_{\max}$, (h) $\varepsilon\rho_S$ (i) $\varepsilon q_{\max}$, (j) $\rho_S q_{\max}$, (k) ρ^2, (l) $q_{\max}^2$.

SA of EM, Fassò et al. GE-2002

Nevertheless, from the first plot of figure above, we see that the conditional variance of $e(\mathbf{x})$ increases with d_p.

Moreover, residuals have high tails and high kurtosis, being $k(e) = 5.07$.

We then fit the skedastic model

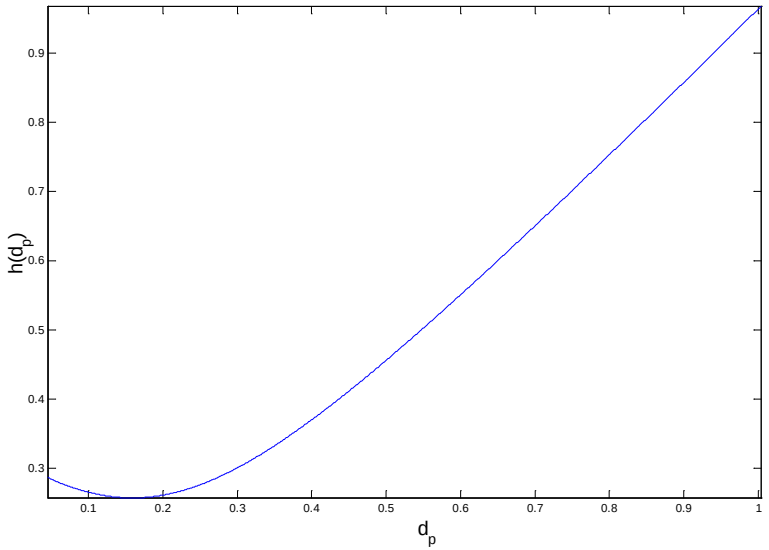
$$h^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2$$

and get the following quadratic equation

$$h(\mathbf{u})^2 = 0.0973(\pm 0.021) - 0.389(\pm 0.093) d_p + 1.221(\pm 0.087) d_p^2$$

Note that, contrary to regression equation, here d_p is the nonzero mean original parameter.

The heteroskedastic model greatly improves error normality since now we have $k\left(\frac{e}{h}\right) = 3.02$.



Skedastic function $h(d_p)$ for LUB

Moreover this figure shows that after fixing ε, ρ_S and $q_{\max}$, output uncertainty strongly depend on d_p being generally increasing with d_p. Hence the heteroskedastic model gives a further reduction of the output uncertainty especially for packed column reactors with small particles diameters d_p.

Multivariate regression

Since the correlation coefficient between the errors of the LUB model and the t_b model 2 is quite small, being -0.109, we can conclude that the residual uncertainty of the two subsystems is almost independent.

Results and Discussion

	t_b	LUB			Multivariate SI
		skedastic			
		SI	SI*	H-SI	
q_{max}	59.75%	34.01%		34.01%	46.88%
ρ_S	36.05%	19.27%		19.27%	27.66%
ε	1.12%	36.47%		36.47%	18.79%
d_p		7.26%	0.61%	7.88%	3.94%
b		1.09%		1.09%	0.54%
$\rho_S - q_{max}$	2.52%	0.15%		0.15%	1.34%
$\varepsilon - q_{max}$	0.21%	0.33%		0.33%	0.27%
$\varepsilon - \rho_S$	0.11%	0.20%		0.20%	0.16%
$d_p - \varepsilon$		0.31%		0.31%	0.15%
$d_p - q_{max}$		0.03%			
Totals	99.77%	99.11%		99.70%	99.74%
R^2	99.77%	99.11%			
Output Std	0.1025	0.060			
Model RMSE	0.015292	0.00023			

Table 2: SA for t_b and LUB.

Technical Details

- The first three lines contain both linear and pure quadratic components.
- The multivariate SI's are based on the simple average formula and the weighted averages are not reported here since, in this case, the results are essentially the same.

Considering the bivariate system altogether:

- the maximum heavy metal up-take (q_{max}) is the most important process parameter

  This conclusion is in agreement with those ones reported in similar works and physically acceptable.
- The biosorbent density (ρ_S) is the second significant parameter.

  As also reported by Hatzikioseyian et al. (2001), the increase in density ρ_S produces an increase in adsorbing material in the fixed bed, improving both t_b and LUB outputs.

Considering the output separately:

- - for LUB: the most relevant factor is the fixed bed void degree (ε) with an influence of 36%:
  - for t_b: q_{max} and ρ_S are the most important .

Summing up

The characteristics of the biosorbent materials seem to be more important for the system altogheter than fluid dynamic factors (in the range of the investigated conditions for the selected equilibrium local model).

This **global result** is also true for the **column working time** t_b alone

but considering the **length of unused bed** LUB the void degree (ε) is quite important.

Conclusions

A standard procedure to evaluate the influence of process parameters on the performance of fixed bed reactors used to remove heavy metals from wastewaters by biosorption has been proposed.

The statistical approach is based on **Extended Sensitivity Analysis** which generalizes standard *SA* to cope with heteroskedastic and multi-output systems.

With this approach, interaction effects can also be estimated.

The influence of the main process parameters on the heavy metal biosorption has been estimated for design purposes and as a useful tool to plan experimental runs in terms of experimental error variance.

Main References

1. Fassò A. 2001. Sensitivity Analysis in A. El-Shaarawi, W. Piegorsch (eds). *Encyclopedia of Environmetrics*. Volume 4, pp 1968-1982, Wiley, New York.

2. A. Fassò, A. Esposito, E. Porcu, A.P. Reverberi, F. Vegliò (2002). Nonlinear Multivariate Statistical Sensitivity Analysis of Environmental Models with Application on Heavy Metals Adsorption from Contaminated Wastewater. Working paper available at http://ingegneria.unibg.it/fasso

3. Saltelli A, Chan K, Scott M. 2000. *Sensitivity Analysis*. Wiley, New York.