Interpretation of non-linear empirical data-based process models using global sensitivity analysis

Tao Chen[a,b,*], Yanhui Yang

[a] *School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore 637459, Singapore*
[b] *Division of Civil, Chemical and Environmental Engineering, University of Surrey, Guildford GU2 7XH, UK*

Abstract

Flexible non-linear regression techniques have been widely used for data-based modeling of chemical processes, and they form the basis of process design under the framework of response surface methodology (RSM). These non-linear models typically achieve more accurate approximation to the factor-response relationship than traditional polynomial regressions. However, non-linear models usually lack a clear interpretation as to how the factors contribute to the prediction of process response.

This paper applies the technique of sensitivity analysis (SA) to facilitate the interpretation of non-linear process models. By recognizing that derivative-based local SA is only valid within the neighborhood of certain "nominal" values, global SA is adopted to study the entire range of the factors. Global SA is based on the decomposition of the model and the variance of response into contributing terms of main effects and interactions. Therefore, the effect of individual factors and their interactions can be both visualized by graphs and quantified by sensitivity indices. The proposed methodology is demonstrated on two catalysis processes where non-linear data-based models have been developed to aid process design. The results indicate that global SA is a powerful tool to reveal the impact of process factors on the response variables.

Keywords: Gaussian process regression, Monte Carlo methods, process modeling, response surface methodology, sensitivity analysis, variance decomposition

1. Introduction

Mathematical models are the foundation of the systems approach to the design of chemical and other processes [1]. Models can be developed through the representation of fundamental principles that govern the process, and thus they are termed *first-principles* or *mechanistic* models. Alternatively, models may be purely based on experimental data and are called *data-based* or *empirical*. A third category is the hybrid modeling, sometimes termed "grey-box modeling", which combines process mechanisms and empirical knowledge from data. Although data-based models are typically reliable only within the operating region where the data are collected, they have seen wide applications due to the simplicity of model development and implementation. This is especially true if the process is still in its early design stage, whereby the time and resources needed for mechanistic modeling are hardly justifiable.

This study is further restricted to batch-wise (as opposed to time-dependent) modeling that relates the process response (y, e.g. product yield) to the operating factors ($\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{T}}$, e.g. reaction temperature and pressure; also termed process variables or input variables): $y = f(\mathbf{x}) + \epsilon$,

*Corresponding author. Tel.: +44 1483 686593, Fax: +44 1483 686581 (T. Chen).
Email address: t.chen@surrey.ac.uk (Tao Chen)

where d is the number of factors and ϵ denotes zero-mean random noise. These models are typically used in off-line design stage to facilitate the understanding and optimization of the process. Such a data-based model is the central component of the so-called response surface methodology (RSM) for rational process design [2, 3, 4, 5].

The traditional method in RSM is to fit a polynomial function (typically linear, quadratic or cubic polynomial) to the experimental data, followed by identifying the process factors that optimize the objective function. A salient advantage of simple polynomial functions is the self-explanatory interpretability of the models. The regression coefficients clearly indicate the sign and magnitude of impact of process factors and factor interactions. Furthermore, a powerful technique, analysis of variance (ANOVA), can be employed to identify whether the impact is statistically significant [5]. In addition, the scope of polynomials may be significantly expanded by applying transformation on the process factors prior to modeling. Non-linear (e.g. logarithmic or logistic) transformation is particularly attractive if it is known *a priori* to result in linear factor-response relationship, and thus linear regression can be used. However in general situation, the prediction accuracy of the polynomial regression is unsatisfactory if the chemical process is complex and does not conform to the restrictive functional form [6, 7, 8, 9]. Consequently, the model-based process understanding and optimization may be unreliable . To address this issue, flexible non-linear models have been applied to provide a more accurate approximation of the process behavior, such as artificial neural network (ANN) [6, 10], support vector machine (SVM) [7], and Gaussian process (GP) regression [1] [9]. Here "flexible" refers to the property that the models are not restricted to a certain functional form, but capable of approximating any function to any accuracy should sufficient data are available. However, these complex models do not allow a clear interpretation with respect to the influence of the process factors on the response. A few researchers have pointed out this issue with flexible non-linear models [11, 12]; yet this topic is still under-explored in the context of data-based process modeling.

This paper proposes to apply the technique of sensitivity analysis (SA) [13] to facilitate the interpretation of data-based non-linear process models. The objective of SA is to understand how changes in the factor $\mathbf{x}$ influence the response y. Two categories of SA are available: local and global analysis. Local SA is based on the partial derivatives $\partial f/\partial x_i, i = 1, \ldots, d$, which need to be calculated at a certain "nominal" point $\mathbf{x}_0$ (e.g. current process operating point). Local SA is a common technique for understanding first-principles chemical models, such as kinetic [14] and thermodynamic models [15], oscillatory biochemical systems [16], among others [13]. Nevertheless, due to the use of derivatives, this approach is only useful within a small neighborhood of $\mathbf{x}_0$ as far as a non-linear model is concerned.

In order to assess the factors' effect within their entire range, global SA is required [17]. Similar to ANOVA, Global SA is to decompose the model and the variance of response into contributing terms of main effects and interactions, and thus factors' influence can be quantified. However, classical ANOVA relies on the simple polynomial parameterization of the model to calculate these effects and interactions, and it is not directly applicable to flexible non-linear process model (e.g. ANN, SVM or GP) [18]. The application of global SA to first-principles chemical models has been well reported [19, 20, 21, 22]. However, to the best knowledge of the authors, using global SA for interpreting data-based process models is largely unexplored in the literature.

1.1. Outline

The rest of this paper is organized as follows. Section 2 outlines the general methodology of data-based modeling for process design. Focus will be given to a specific modeling approach, Gaussian process (GP) regression. Note that the global SA approach is equally applicable to other model structures, such as ANN and SVM. Section 3 discusses the formulation and computation of

[1] Note that the term "process" in GP refers to stochastic process in a mathematical sense, whilst in process modeling it refers to real chemical or manufacturing process. By consulting the context, these two "processes" should not be confused.

SA methods. Two examples, related to the design of catalytic reaction systems, will be presented in Section 4 to demonstrate how global SA is useful in interpreting the process models. Finally Section 5 concludes this paper.

2. Data-based modeling to aid process design

RSM is an important approach to data-based modeling for process design. RSM mainly comprises three components: (i) design of experiments (DoE) to determine the process factors' values based on which experiments are conducted and data are collected; (ii) empirical modeling to approximate the factor-response relationship (i.e. the response surface); and (iii) optimization to find the best response value based on the empirical model. The classical DoE methods, such as factorial design, typically assign two or three pre-determined levels for each process factor, and experiments are conducted at the combination of the levels of different factors. Using a small number of levels is appealing if the factors' values are difficult to change in practice. However, this strategy may not give an optimal coverage of the design space due to limited levels of the factors being studied, and thus it may result in a less reliable empirical model [23]. An alternative approach is the "space-filling" designs that allocate design points to be uniformly distributed within the range of each factor [23, 24]. Latin hypercube sampling (LHS) [24] and uniform design [23] are among the most widely used space-filling designs in practice.

After conducting experiments at the designed points, the data are traditionally modeled by the following polynomial regression:

$$f(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{1 \le i \le j \le d} w_{i,j} x_i x_j + \cdots + w_{1,\ldots,d} \prod_{i=1}^{d} x_i \tag{1}$$

where w's are the regression coefficients and they can be estimated by using least-squares given a set of experimental data. Depending on applications, the polynomial regression is typically restricted to the first, second (quadratic) or third (cubic) order. For a quadratic regression, a total of $d(d+1)/2+1$ coefficients need to be estimated, and this number may be higher than the number of data points collected from experiments. Hence, the least-squares solution becomes ill-conditioned. This problem can be resolved by using ridge regression [25], partial least squares [26], or stepwise variable selection [27].

The major issue with polynomial regression is the prediction accuracy. Since the model is restricted to the polynomial form, it is not capable of approximating complex factor-response relationship. To address this issue, more flexible data-based models have been adopted, including ANN [6, 10], SVM [7] and GP regression [9]. Unlike polynomial regression, these complex models do not allow a transparent interpretation as to which process factors contribute the most to the response variable. In this study, the GP model is used as testbed for illustrating the use of global SA for model interpretation.

2.1. The GP regression model

GP, also termed kriging model in the literature with slightly different formulation [28], is a flexible modeling technique that can be used for both regression and classification purposes [29]. The GP regression model can be derived from the perspectives of ANN and Bayesian non-parametric regression; see [29] for details. In recent studies, GP model has been shown to give superior prediction accuracy in process control [30], chemometric calibration [31], medical treatment design [32], and RSM [8, 9]. Some theoretical analysis of the GP's generalization performance, which measures the prediction capability on unseen data, is given in [29, Chapter 7]. In this subsection, a brief overview of Gaussian process regression model is given, including three components: (i) the probabilistic formulation and parameterization of the model (eqs. (2)(3)), (ii) the prediction formulae (eqs. (4)(5)), and the maximum likelihood method for parameter estimation (eq. (6)).

Consider a data set consisting of n data points, $\{\mathbf{x}^{(i)}, y^{(i)}; i = 1, \ldots, n\}$, where we use super-script to index data points and sub-script to index dimensions. A GP for regression is defined such that the regression function $y(\mathbf{x})$ has a Gaussian prior distribution with zero mean, or in discrete form:

$$\mathbf{y} = (y^{(1)}, \ldots, y^{(n)})^{\mathrm{T}} \sim G(\mathbf{0}, \mathbf{C}) \tag{2}$$

where $\mathbf{C}$ is an $n \times n$ covariance matrix of which the ij-th element is defined by the covariance function: $C(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. An example of such a covariance function is:

$$C(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = a_0 + a_1 \sum_{k=1}^{d} x_k^{(i)} x_k^{(j)} + v_0 \exp\left(-\sum_{k=1}^{d} w_k (x_k^{(i)} - x_k^{(j)})^2 \right) + \delta_{ij}\sigma^2 \tag{3}$$

where $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^{\mathrm{T}}$; $\delta_{ij} = 1$ if $i = j$, otherwise it is equal to zero. We denote $\boldsymbol{\theta} = (a_0, a_1, v_0, w_1, \cdots, w_d, \sigma^2)^{\mathrm{T}}$ as "hyper-parameters" defining the covariance function. The hyper-parameters must be non-negative to ensure that the covariance matrix is non-negative definite. For the covariance function depicted in eq. (3), the first two terms represent a constant bias (offset) and a linear correlation term, respectively. The exponential term is similar to the form of a radial basis function, and it takes into account the potentially strong correlation between the outputs for nearby inputs. The term σ^2 captures the random error effect. By combining both linear and non-linear terms in the covariance function, GP is capable of handling both linear and non-linear data structures [31]. Other forms of covariance functions are discussed in [29]. Note that unlike classical polynomials, the number of hyper-parameters in a GP is not directly linked to the complexity of the model. A thorough discussion on model complexity through a Bayesian perspective is given in [33].

The prediction at a new data point $\mathbf{x}^*$ is also Gaussian distributed: $y^* \sim G(\hat{y}^*, \sigma^2_{\hat{y}^*})$ with

$$\hat{y}^* = \mathbf{k}^{\mathrm{T}}(\mathbf{x}^*)\, \mathbf{C}^{-1}\mathbf{y} \tag{4}$$

$$\sigma^2_{\hat{y}^*} = C(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^{\mathrm{T}}(\mathbf{x}^*)\, \mathbf{C}^{-1}\, \mathbf{k}(\mathbf{x}^*) \tag{5}$$

where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}^*, \mathbf{x}^{(1)}), \ldots, C(\mathbf{x}^*, \mathbf{x}^{(n)})]^{\mathrm{T}}$ and the covariance functions are calculated from eq. (3). The capability to providing the prediction uncertainty in terms of the variance is an important feature of GP for robust process design [9, 8]. In this study, we will focus on interpreting how the mean prediction is affected by process factors.

The hyper-parameters $\boldsymbol{\theta}$ can be estimated by maximizing the following log-likelihood function using optimization algorithms:

$$L = -\frac{1}{2}\log\det\mathbf{C} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y} - \frac{n}{2}\log 2\pi \tag{6}$$

A conjugate gradient method is usually used to find the hyper-parameters that maximize the above likelihood [29]. A Matlab implementation of the GP model is publicly available [34], and it was used to produce the results in this study.

It should also be noted that the calculation of the likelihood involves a matrix inversion step and takes time of the order $O(N^3)$, which can be extremely demanding for large data set. Fortunately, in the context of model-based process design, the experiments are costly to run, and the available data are normally limited and should not pose a computational problem for GP modeling.

3. Model interpretation using sensitivity analysis

3.1. Local sensitivity analysis

The local sensitivity analysis is based on a direct application of differentiation: $D_i = \partial f / \partial x_i$ [13]. Given a specific process model, the derivatives can be calculated by either analytical method

or finite difference. The quantity D_i indicates how sensitive the response is to a perturbation of the i-th factor. It is a local measure since the derivative needs to be evaluated at a "nominal" operating condition; that is, D_i can only be calculated with respect to a fixed point of the factors. Therefore, local SA is only valid at a small neighborhood of the nominal point, unless the process model is a simple linear function. It is not suitable to tell how the response variable behaves when the factors vary across the whole design space. More seriously, local SA is a one-factor-at-a-time approach, since the sensitivity of the i-th factor is obtained by assuming all other factors are fixed. This approach contradicts the fundamental principle of RSM, which aims at investigating the effect of both individual factors and their interactions [5].

3.2. Global sensitivity analysis

Global SA is based on a decomposition of the model $f(\mathbf{x})$ into main effects and interactions [17]. The main effects quantify the impact of individual process factors, $x_i, i = 1, \ldots, d$, whilst the interactions correspond to the joint effects of multiple factors. In particular,

$$\begin{aligned} f(\mathbf{x}) =& E(y) + \sum_{i=1}^{d} z_i(x_i) + \sum_{1\le i<j\le d} z_{i,j}(\mathbf{x}_{i,j}) \\ &+ \sum_{1\le i<j<k\le d} z_{i,j,k}(\mathbf{x}_{i,j,k}) + \ldots + z_{1,2,\ldots,d}(\mathbf{x}) \end{aligned} \tag{7}$$

where $E(\cdot)$ denotes expectation and $E(y)$ is the overall mean of the model. The main effects are

$$z_i(x_i) = E(y|x_i) - E(y) \tag{8}$$

and the two-factor interactions are

$$z_{i,j}(\mathbf{x}_{i,j}) = E(y|\mathbf{x}_{i,j}) - z_i(x_i) - z_j(x_j) - E(y) \tag{9}$$

Similar expressions can be derived for multi-factor interactions. Global SA is also termed probabilistic SA, because the expectation operations are carried out with respect to a certain probability density function (*pdf*) of process factors [35]. For the analysis of parameters (such as reaction rate constant) in a first-principles model, these parameters are treated as factors and are typically given a normal distribution. In the context of interpreting data-based process models, process factors are instead given a uniform distribution within a certain range.

Note that the main effects and interactions are all functions of process factors. Therefore, computing and plotting the main effects and/or interactions against the factors is a powerful graphic tool to depict how the process responds to the factors and their interactions. For example, if x_i is an important factor to influence the response, then the conditional expectation $E(y|x_i)$ will have a large variation across x_i values. This observation motivated the following variance-based measure to quantify the importance of factor x_i [17]:

$$V_i = \mathrm{var}\{z_i(x_i)\} = \mathrm{var}\{E(y|x_i)\} \tag{10}$$

and interactions

$$V_{i,j} = \mathrm{var}\{z_{i,j}(\mathbf{x}_{i,j})\} = \mathrm{var}\{E(y|\mathbf{x}_{i,j})\} - V_i - V_j \tag{11}$$

In the context of process design, the data are usually obtained from designed experiments. Typical DoEs, be it factorial design or Latin hypercube sampling, will ensure that process factors are independent. Under this assumption, the total variance of y, $V = \mathrm{var}(y)$, can be decomposed as

$$V = \sum_{i=1}^{d} V_i + \sum_{1 \le i < j \le d} V_{i,j} + \sum_{1 \le i < j < k \le d} V_{i,j,k} + \ldots + V_{1,2,\ldots,d} \tag{12}$$

and thus the sensitivity indices are normalized as $S_i = V_i/V$, $S_{i,j} = V_{i,j}/V$, and so on.

Another useful measure is the variance of *total effect* of the i-th factor defined by

$$V_{Ti} = V - \text{var}\{E(y|\mathbf{x}_{-i})\} \tag{13}$$

where $\mathbf{x}_{-i}$ denotes the sub-vector of $\mathbf{x}$ containing all elements except x_i. V_{Ti} measures the variance of y that remains if the true values of $\mathbf{x}_{-i}$ can be determined. The corresponding total sensitivity index is then

$$S_{Ti} = V_{Ti}/V \tag{14}$$

The total effect index accounts for the total contribution to the response variation due to factor x_i, including its first-order effect S_i plus all higher-order effects due to interactions. As such, $S_{Ti} \geq S_i$, where equality holds when x_i is not interacting with any other factors.

3.3. Computation for global SA using Monte Carlo methods

In this sub-section, the computational strategy for global SA is discussed, including the calculation of main effects and interactions as in eq. (7) and the sensitivity indices S_i and S_{Ti}. The number of terms that need to be computed increases exponentially with the number of process factors. To alleviate the computational cost, the usual approach is to assess the main effects first. If two-factor interactions are deemed to be significant, then they should also be calculated. Finally, higher-order effects typically have small magnitude and thus are usually not considered.

The computation for global SA mainly involves the evaluation of the following integral:

$$E(y|x_i) = \int y(\mathbf{x}) p(\mathbf{x}_{-i}|x_i) d\mathbf{x}_{-i} \tag{15}$$

and its variance with respect to x_i: $\text{var}\{E(y|x_i)\}$, which is also an integral by definition. Here $p(\mathbf{x}_{-i}|x_i)$ is the condition *pdf* of $\mathbf{x}_{-i}$ given x_i, which can be derived from the overall *pdf* $p(\mathbf{x})$. In some rare cases these quantities can be obtained analytically, for example when the process model $y(\mathbf{x})$ is a Gaussian process and the factors $\mathbf{x}$ is normally distributed [35]. When $\mathbf{x}$ is uniformly distributed as in the case of model interpretation, analytical solutions are not available. We adopt the Monte Carlo (MC) method that is widely used for global SA to compute these quantities [17]. The basic concept of the MC method is given below; more details are available from [17].

The principle of MC methods is to generate a large number of random samples that are distributed according to a certain *pdf*, and then any integral with respect to this distribution can be approximated by using these MC samples. For example, to calculate the expectation in eq. (15), one may draw q random samples from $p(\mathbf{x})$: $\mathbf{x}^{(k)}, k = 1, \ldots, q$. Here $p(\mathbf{x})$ is an independent uniform distribution, i.e. $p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i)$ where $p(x_i)$ is uniformly distributed within the specified range for the i-th factor. The range of the i-th factor can be equally divided into g intervals: $[r_i^{(0)}, r_i^{(1)}, \ldots, r_i^{(g)}]$. Furthermore, note that $E(y|x_i)$ is a function of x_i. Then, the expectation can be calculated for each interval $I_i^{(h)} = [r_i^{(h-1)}, r_i^{(h)}]$ as:

$$E(y|r_i^{(h-1)} \leq x_i < r_i^{(h)}) \approx \frac{1}{||I_i^{(h)}||} \sum_{k: x_i^{(k)} \in I_i^{(h)}} y\left(\mathbf{x}_{-i}^{(k)}, x_i^{(k)}\right) \tag{16}$$

where the MC samples are expressed as $\mathbf{x}^{(k)} = [\mathbf{x}_{-i}^{(k)}, x_i^{(k)}]$, and $||I_i^{(h)}||$ denotes the number of random samples whose i-th element falls within the interval $I_i^{(h)}$. Overall, the process model needs to be evaluated q times at all the random samples (i.e. making prediction for q times). Based

on the approximation in eq. (16), the variance var$\{E(y|x_i)\}$ can be obtained to calculate the sensitivity index for the i-th factor. Similar method can be developed for interaction terms.

This basic MC method requires to determine two parameters, the number of random samples q and the number of intervals g. To reach an accurate MC approximation, q can vary from several thousands to tens of thousands, and a suitable number is largely dependent on the smoothness of the response surface and the number of factors [17]. The number of intervals should be selected such that a sufficient number of random samples (e.g. 100) fall within each interval for reliable estimation as in eq. (16).

The major computation in the MC method is to evaluate the non-linear process model for q times. Although the model has a "complex" non-linear form, its evaluation is usually very quick in comparison with evaluating a first-principles model. Nevertheless, it has been realized that structured or even deterministic sampling can provide the same order of accuracy with a dramatically smaller number of samples. These sampling methods include Latin hypercube sampling [24], Hammersley sequence sampling [36], and uniform design [23]. Discussions on more efficient algorithms by exploring the properties of sensitivity index are given in [37].

3.4. Other related computation methods

In addition, it is worth noting that an alternative approach for computing global SA is based on Fourier transforms and named Fourier amplitude sensitivity test (FAST) [19, 38]. FAST decomposes the variance of process response using spectral analysis, and then assesses the influence of individual factors by its contribution to the total variance. For researchers who are not interested in the detailed computational methods, both MC and FAST methods have been implemented in the dedicated SA software SIMLAB [17], which is publicly available [39]. SIMLAB was used as a Matlab toolbox to produce the results presented in this paper; it may also be used as a standalone package.

4. Case study

This section demonstrates the application of global SA to interpret the GP models that are developed to aid the understanding and design of two catalytic processes: the oxidation of benzyl alcohol [27] and the epoxidation of *trans*-stilbene [8].

4.1. Case 1: the benzyl alcohol oxidation process

Oxidation of alcohols into the corresponding aldehydes or ketones, in particular benzyl alcohol to benzaldehyde, is one of the most important functional group transformations in organic synthesis. Experiments were conducted to study the impact of various process factors on the conversion of benzyl alcohol [27]. The selected catalyst, K-Mn/C, was prepared by co-impregnating aqueous solutions of potassium and manganese nitrates onto commercially available activated carbon. The catalytic oxidation process was conducted in a bath-type lab-scale reactor. More experimental details can be found in [27]. The conversion of the raw material is regarded as the process response variable and is calculated on the basis of moles of benzyl alcohol as follows:

$$\text{Conversion}(\%) = \frac{(\text{initial moles}) - (\text{final moles})}{(\text{initial moles})} \times 100\% \tag{17}$$

The analytical procedure to determine the concentrations was reported elsewhere [27]. Repeated experiments indicated that the standard deviation of conversion is typically within 2%. Five process factors are considered, including reaction temperature, partial pressure of oxygen, concentration of benzyl alcohol (in terms of mmol diluted within 10 ml of the solvent, toluene), percentage of Mn, and K:Mn ratio. The range of these factors is listed in Table 1.

(Table 1 about here)

The original purpose of the experiments was to develop a quadratic regression model to relate the conversion to the five process factors. Hence, the central composite design, which is especially appropriate for quadratic regression [5], was adopted to give 32 experimental runs. In a later stage, an additional eight experiments were conducted to further confirm the effect of increasing K:Mn ratio. Therefore, the data set is not the result of rigorously designed experiments, which is not uncommon in practical experimentation. Nevertheless, our main focus is to demonstrate the use of global GA for model interpretation, as opposed to discussing the most appropriate DoE method. The data for the 38 experimental runs have been published in [27].

Next, the modeling performance of the polynomial (specifically the quadratic model with ridge regression) and GP models is briefly compared, since it is not the primary objective of this paper. For this purpose, we adopt the method of leave-one-out cross-validation (LOOCV) [40]. LOOCV takes a single data point from the entire data set as the validation data, and then develop a model using the remaining data points. Hence the error for the validation data can be calculated. This procedure is repeated such that each data point is used once for validation, and the overall validation error (usually in terms of root mean squared error ($\mathrm{RMSE_{cv}}$) and coefficient of determination (R^2_{cv})) is used as the criterion to assess model quality. These two metrics on the modeling data, denoted by $\mathrm{RMSE_m}$ and R^2_{m} respectively, are also shown. The results are summarized in Table 2. It appears that the achieved $\mathrm{RMSE_m}$ for the two models is in line with the standard deviation of the measured conversion. However, the within-model performance is significantly better than the LOOCV results, suggesting that both quadratic and GP models overfit the data. The overfitting problem may also result from limited experimental data that do not sufficiently cover the entire factors' space. In this situation, LOOCV requires the model to be extrapolated to under-explored region, giving rise to large errors. Nevertheless, the results clearly favor the GP model since it attains lower prediction error than quadratic regression. The LOOCV results are also compared in Figure 1.

(Table 2 and Figure 1 about here)

Even if quadratic regression provides clear information as to how the factor affect the *modeled* response, this interpretation is unreliable due to the large mismatch between *modeled* and *actual* responses. In addition, the relatively low R^2 value for both models (compared with the other case study presented in the next sub-section) may be the result of the flawed DoE. It is well recognized that if the data do not provide good coverage of the design space, then the data-based model will not achieve excellent accuracy. Another possibility is that the conversion is significantly affected by other factors, which were not considered or not well-controlled when conducting experiments.

The global sensitivity indices are calculated based on 10,000 MC samples and are listed in Table 1. Evaluating the GP model for 10,000 times took 41.9 s for this case. (All computation reported in this paper was conducted under Matlab environment on a Pentium 3.4 GHz computer running Windows XP.) According to the main effect indices S_i, reaction temperature (x_1) has the most influence on conversion, followed by Mn loading (x_4) and K:Mn ratio (x_5). The other two factors have relatively small impact. In addition, the total indices S_{Ti} are very similar to the main effect indices, which means that the impact of interaction terms is negligible. The same conclusion can be reached by observing that the sum of S_i's is close to one, which suggests that the main effects have accounted for the majority of the variance in the response variable.

Figure 2 displays the main effects $E(y|x_i)$ as a function of x_i (scaled to $[0, 1]$) within the range of each process factor, whereby the range is divided into ten intervals. Except for initial concentration of benzyl alcohol (x_3), the other four factors have positive effect on the conversion, i.e. an increase of factor's value leads to an increase of conversion. Again, the effect of reaction temperature is the most important. According to thermodynamics, for an endothermic reaction like benzyl alcohol oxidation, higher temperature results in higher equilibrium constant and thus higher conversion. An increase in the partial pressure of oxygen will initially increase the conversion, and then has minor impact once it reaches a certain level. The impact of Mn loading and K:Mn ratio has similar

patterns. Finally, initial concentration of benzyl alcohol has an negative impact on conversion, although this impact is not as significant as other factors according to Table 1.

(Figure 2 about here)

It appears that to further improve the conversion, one should increase all factors but initial concentration of benzyl alcohol. In practice, a more rigorous approach is to optimize the conversion based on the GP-model, and this topic has been discussed elsewhere [8]. When conducting process optimization, the constraints on process factors must also be considered, For example, if the reaction temperature is increased over the limit for carbon oxidation, it will damage the catalyst support (activated carbon) and thus have adverse impact on benzyl alcohol oxidation. Under these situations, constrained optimization methods, such as sequential quadratic programming, should be used.

4.2. Case 2: the trans-stilbene epoxidation process

The second example is to study a catalytic epoxidation process that converts *trans*-stilbene into stilbene oxide using molecular oxygen as the oxidant. Stilbene oxide is a commercially important intermediate used in the synthesis of various fine chemicals and pharmaceuticals. Experiments were conducted to test the effectiveness of a novel heterogeneous catalyst: cobalt ion-exchanged faujasite zeolite (Co^{2+}-NaX) [8]. The experimental protocol to prepare Co^{2+}-NaX catalyst is described elsewhere [8]. The liquid-phase catalytic *trans*-stilbene epoxidation reactions were carried out using a batch-type reactor operated under atmospheric pressure. After reaction, the solid catalyst was filtered off, and the liquid organic products were analyzed by an Agilent gas chromatograph (GC) 6890. The conversion is regarded as the process response and is calculated on the basis of moles of *trans*-stilbene, similar to eq. (17). More detailed information for the analytical procedure is given in [8], where the standard deviation of the measured conversion was reported to be within 1%. Five process factors were considered, including reaction temperature, partial pressure of oxygen, initial *trans*-stilbene concentration, stirring rate and reaction time. The range of these factors to be explored is listed in Table 3.

(Table 3 about here)

The original purpose of this experimental study was to develop a GP-based iterative optimization method to maximize the *trans*-stilbene conversion [8]. In order to facilitate the process design, an incremental Latin hypercube sampling design was developed to decide the factor's values for experimentation. A total of 41 experimental runs, corresponding to 41 data points, were conducted iteratively, whilst in each iteration a GP model was developed to aid process optimization. In current study, a single GP model is developed using all the data to demonstrate how global SA is useful to help interpret the model.

A brief comparison between the GP regression and the traditional quadratic model with ridge regression is given in Table 4, which lists the RMSE and R^2 on the modeling and LOOCV data. The prediction performance is also illustrated in Figure 3 based on the LOOCV procedure. Note that the LOOCV results are slightly different from those presented in [8], in which the models were based on 40 data points. The achieved $\mathrm{RMSE_m}$ for the two models appears to be consistent with the standard deviation of the measured conversion. In addition, the difference between within-model and LOOCV performance is less than that for the benzyl alcohol example, suggesting less degree of overfitting for both models. In terms of LOOCV, both methods achieve higher prediction accuracy than they did on the benzyl alcohol data. This may be attributable to the use of LHS design that gives better coverage of the design space. Again, GP model ($\mathrm{RMSE_{cv}}$=4.50, R^2_{cv}=0.96) is preferred to quadratic regression ($\mathrm{RMSE_{cv}}$=5.88, R^2_{cv}=0.92) in terms of more accurate approximation to the response-factor relationship.

(Table 4 and Figure 3 about here)

An additional note from Figure 3 is that all data but one (93.5% conversion) give a conversion lower than 70%. This high conversion may be an "outlier" and excluded from modelling and interpretation. By doing so, the need for extrapolation to high conversion region is removed and thus the modelling performance would be improved. However, the high conversion was the result of process optimization [8] and is a *genuine* data point. Modelling with such an "outlier" excluded does not fully utilize all available information of this process. Certainly, a better practice is to collect more data whose conversion is between 70% and 93.5% to improve the model; yet it is not usually done in practice if the (near-)optimal conversion has already been identified. From another perspective, extrapolation is unreliable in prediction but useful in optimization. There is a balance with regard to experimenting at know region (interpolation to improve model) and exploring unknown region (extrapolation in the hope to find better response). This topic was examined in detail in [41].

The global sensitivity indices for the five factors are calculated based on 10,000 MC samples and are listed in Table 3. Evaluating the GP model for 10,000 times took 43.2 s for this case. The main effects are also graphically illustrated in Figure 4. Similar to the benzyl alcohol oxidation, the *trans*-stilbene epoxidation is also an endothermic reaction, and thus higher temperature is preferred for higher conversion. This is confirmed by both a large sensitivity index $S_1 = 0.445$ and the dominant impact of temperature on conversion as given in Figure 4. The reaction time x_5 has the second highest impact, and the graph of $E(y|x_5)$ favors a longer reaction time. For this reaction, it appears that the oxygen pressure (x_2) and stirring rate (x_4) do not significantly affect conversion, and thus they may not need to be considered in further studies. Similar to benzyl alcohol oxidation, the initial concentration of raw material, *trans*-stilbene, has an negative impact on conversion.

(Figure 4 about here)

The results from global SA also indicates that the impact from interaction terms is not negligible. In Table 3, there is appreciable difference between the main (S_i) and total (S_{Ti}) indices, in particular for temperature and reaction time. In addition, the sum of five main indices is 0.752, suggesting that the interactions would account for $1 - 0.752 = 24.8\%$ of the total variation in the response variable. Therefore, further analysis is required to assess the contribution from interaction terms. Table 5 gives the two-factor interaction indices. Except for the interaction between reaction temperature and time ($S_{1,5} = 0.167$), the other items are insignificant. The summation of main effect and two-factor interaction indices is 0.992, which implies that investigation on higher-order interactions is not necessary. To further illustrate the interacting effect of reaction temperature and time, Figure 5 depicts $E(y|x_1, x_5)$ as a function of these two factors (range scaled to $[0, 1]$). For comparison, the same graph for temperature and stirring rate where the interaction is small ($S_{1,4} = 0.006$) is given in Figure 6. In Figure 5, when temperature is low, varying reaction time does not significantly affect conversion; when temperature is high, increasing reaction time is clearly advantageous. The graph indicates a clear positive interaction between these two factors. In contrast, Figure 6 shows that temperature has large impact on conversion no matter what stirring rate is, and varying stirring rate has low influence no matter what temperature is. Hence, it can be concluded from the figure that the interaction between temperature and stirring rate is negligible.

(Table 5, Figure 5 and Figure 6 about here)

5. Conclusions

Data-based modeling with advanced complex regression techniques has been widely used to aid the design and development of chemical and other processes. Compared with traditional polynomial regression, these complex models typically attain more accurate approximation to the underlying process, yet they are more difficult to interpret as to how the modeled process

response is affected by the factors. This paper has demonstrated the application of global sensitivity analysis that provides valuable insight into these models. This approach gives both quantitative assessment of the relative importance of factors and powerful graphic tool to visualize factors' impact. Two examples have been presented to illustrate the effectiveness of this method. For the first example (oxidation of benzyl alcohol), the results suggested that the reaction temperature and Mn loading in the catalyst played paramount impact on the response variable (the conversion of benzyl alcohol), and the five process factors affected the response largely independently with negligible interactions. For epoxidation of *trans*-stilbene, global sensitivity analysis also indicated the reaction temperature as the most important factor, as well as the significance of interaction between reaction temperature and reaction time. Together with existing chemical knowledge, these findings facilitated the understanding of the processes under development, and are useful to guide process optimization in the future.

References

[1] K. Klatt, W. Marquardt, Perspectives for process systems engineeringPersonal views from academia and industry, Computers and Chemical Engineering 33 (3) (2009) 536–550.

[2] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, S. Wold, Design of Experiments: Principles and Applications, Umetrics Academy, 2000.

[3] K. Esbensen, Multivariate Data Analysis: In Practice: An Introduction to Multivariate Data Analysis and Experimental Design, 5th Edition, Camo Process AS, 2002.

[4] T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, A. Nyström, J. Pettersen, R. Bergman, Experimental design and optimization, Chemometrics and Intelligent Laboratory Systems 42 (1998) 3–40.

[5] R. H. Myers, D. C. Montgomery, Response Surface Methodology, Wiley, 1995.

[6] J. R. Dutta, P. K. Dutta, R. Banerjee, Optimization of culture parameters for extracellular protease production from a newly isolated pseudomonas sp. using response surface and artificial neural network models, Process Biochemistry 39 (2004) 2193–2198.

[7] M. Hadjmohammadi, K. Kamel, Response surface methodology and support vector machine for the optimization of separation in linear gradient elution, Journal of Separation Science 31 (2008) 3864–3870.

[8] Q. Tang, Y. Lau, S. Hu, W. Yan, Y. Yang, T. Chen, Response surface methodology using Gaussian processes: towards optimizing the trans-stilbene epoxidation over Co^{2+}-NaX catalysts, Chemical Engineering Journal 156 (2010) 423–431.

[9] J. Yuan, K. Wang, T. Yu, M. Fang, Reliable multi-objective optimization of high-speed WEDM process based on Gaussian process regression, International Journal of Machine Tools and Manufacture 48 (2008) 47–60.

[10] P. Shao, S. T. Jiang, Y. J. Ying, Optimization of molecular distillation for recovery of tocopherol from rapeseed oil deodorizer distillate using response surface and artificial neural network models, Food and Bioproducts Processing 85 (2007) 85–92.

[11] P. Brown, Discussion of the paper by Chen, Morris and Martin, Chemometrics and Intelligent Laboratory Systems 87 (1) (2007) 94–95.

[12] T. Plate, Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using Gaussian process models, Behaviormetrika 26 (1) (1999) 29–50.

[13] A. Saltelli, M. Ratto, S. Tarantola, F. Campolongo, Sensitivity analysis for chemical models, Chemical Reviews 105 (7) (2005) 2811–2827.

[14] T. Turányi, Sensitivity analysis of complex kinetic systems. Tools and applications, Journal of Mathematical Chemistry 5 (3) (1990) 203–248.

[15] J. Bromly, F. Barnes, S. Muris, X. You, B. Haynes, Kinetic and thermodynamic sensitivity analysis of the NO-sensitised oxidation of methane, Combustion Science and Technology 115 (4) (1996) 259–296.

[16] D. Zak, J. Stelling, F. Doyle, Sensitivity analysis of oscillatory (bio) chemical systems, Computers and Chemical Engineering 29 (3) (2005) 663–673.

[17] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global Sensitivity Analysis: The Primer, Wiley-Interscience, 2008.

[18] W. Chen, R. Jin, A. Sudjianto, Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty, Journal of Mechanical Design 127 (2005) 875–886.

[19] G. McRae, J. Tilden, J. Seinfeld, Global sensitivity analysis–a computational implementation of the Fourier amplitude sensitivity test (FAST), Computers and Chemical Engineering 6 (1) (1982) 15–25.

[20] S. Chhatre, R. Francis, A. Newcombe, Y. Zhou, N. Titchener-Hooker, J. King, E. Keshavarz-Moore, Global sensitivity analysis for the determination of parameter importance in the chromatographic purification of polyclonal antibodies, Journal of Chemical Technology and Biotechnology 83 (2008) 201–208.

[21] M. Degerman, K. Westerberg, B. Nilsson, A model-based approach to determine the design space of preparative chromatography, Chemical Engineering and Technology 32 (2009) 1195–1202.

[22] M. Haaker, P. Verheijen, Local and global sensitivity analysis for a reactor design with parameter uncertainty, Chemical Engineering Research and Design 82 (5) (2004) 591–598.

Table 1: Process factors considered to study the benzyl alcohol oxidation process. The sensitivity indices were based on Global SA for the GP model of this process. The sum of sensitivity indices S_i is 0.971.

Process factor	Range of values	S_i	S_{Ti}
Temperature, x_1 (°C)	$60 - 110$	0.390	0.408
Partial pressure of oxygen, x_2 (Bar)	$0 - 1$	0.114	0.119
Concentration of benzyl alcohol, x_3 (mmol/10 mL)	$1 - 4$	0.097	0.104
Manganese loading, x_4 (%)	$1 - 15$	0.206	0.225
K:Mn ratio, x_5	$0 - 4$	0.164	0.176

Table 2: Comparison of modeling performance for the benzyl alcohol oxidation process.

	$\mathrm{RMSE_m}$	$\mathrm{RMSE_{cv}}$	R^2_{m}	R^2_{cv}
Quad.	3.31	12.35	0.97	0.59
GP	1.32	8.36	0.98	0.81

[23] K. T. Fang, D. K. J. Lin, P. Winker, Y. Zhang, Uniform design: theory and application, Technometrics 42 (2000) 237–248.

[24] M. D. McKay, B. J. Beckman, W. J. Conover, A comparison of three methods for selecting values for input variables in the analysis of output from a computer code, Technometrics 21 (1979) 239–245.

[25] E. Vigneau, M. F. Devaux, E. M. Qannari, P. Robert, Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration, Journal of Chemometrics 11 (1997) 239–249.

[26] P. Geladi, B. R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chimica Acta 185 (1986) 1–17.

[27] Q. Tang, Y. Chen, C. Zhou, T. Chen, Y. Yang, Statistical modelling and analysis of the aerobic oxidation of benzyl alcohol over K–Mn/C catalysts, Catalysis Letters 128 (2009) 210–220.

[28] J. Sacks, W. Welch, T. Mitchell, H. Wynn, Design and analysis of computer experiments, Statistical Science 4 (1989) 409–423.

[29] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[30] B. Likar, J. Kocijan, Predictive control of a gas-liquid separation plant based on a Gaussian process model, Computers and Chemical Engineering 31 (2007) 142–152.

[31] T. Chen, J. Morris, E. Martin, Gaussian process regression for multivariate spectroscopic calibration, Chemometrics and Intelligent Laboratory Systems 87 (2007) 59–67.

[32] R. Kamnik, J. Shi, R. Murray-Smith, T. Bajd, Nonlinear modeling of FES-supported standing-up in paraplegia for selection of feedback sensors, IEEE Transactions on Neural Systems and Rehabilitation Engineering 13 (2005) 40–52.

[33] D. J. C. MacKay, Bayesian interpolation, Neural Computation 4 (1992) 415–447.

[34] `http://www.gaussianprocess.org/gpml/code/matlab/doc/`, last accessed on 31 January 2011.

[35] J. Oakley, A. O'Hagan, Probabilistic sensitivity analysis of complex models: a Bayesian approach, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 66 (3) (2004) 751–769.

[36] J. Kalagnanam, U. Diwekar, An efficient sampling technique for off-line quality control, Technometrics 39 (1997) 308–319.

[37] A. Saltelli, Making best use of model valuations to compute sensitivity indices, Computer Physics Communications 145 (2002) 280–297.

[38] R. Cukier, H. Levine, K. Shuler, Nonlinear sensitivity analysis of multiparameter model systems, Journal of Physical Chemistry 81 (1977) 2365–2366.

[39] `http://simlab.jrc.ec.europa.eu/`, last accessed on 31 January 2011.

[40] H. Martens, P. Dardenne, Validation and verification of regression in small data sets, Chemometrics and intelligent laboratory systems 44 (1-2) (1998) 99–121.

[41] S. Valero, E. Argente, V. Botti, J. Serra, P. Serna, M. Moliner, A. Corma, DoE framework for catalyst development based on soft computing techniques, Computers and Chemical Engineering 33 (2009) 225–238.

Table 3: Process factors considered to study the *trans*-stilbene epoxidation process. The sensitivity indices were based on Global SA for the GP model of this process. The sum of sensitivity indices S_i is 0.752.

Process factor	Range of values	S_i	S_{Ti}
Temperature, x_1 (°C)	$60 - 120$	0.445	0.683
Partial pressure of oxygen, x_2 (Bar)	$0.2 - 0.8$	0.013	0.030
Initial stilbene concentration, x_3 (mmol/15 mL)	$1 - 5$	0.125	0.192
Stirring rate, x_4 (rpm)	$200 - 1250$	0.010	0.021
Reaction time, x_5 (min)	$30 - 240$	0.159	0.357

Table 4: Comparison of modeling performance for the *trans*-stilbene epoxidation process.

	RMSE_m	RMSE_cv	R^2_m	R^2_cv
Quad.	2.53	5.88	0.98	0.92
GP	1.54	4.50	0.99	0.96

Table 5: Two-factor interaction indices for the GP model of the *trans*-stilbene epoxidation process. The summation of main effect and two-factor interaction indices, i.e. $\sum_{i=1}^{5} S_i + \sum_{1 \le i < j \le 5} S_{i,j}$, is 0.992.

$S_{i,j}$	Temperature, x_1	Pressure, x_2	Concentration, x_3	Stirring, x_4
Pressure, x_2	0.013			
Concentration, x_3	0.028	0.011		
Stirring, x_4	0.006	0.000	0.000	
Time, x_5	**0.167**	0.004	0.009	0.002

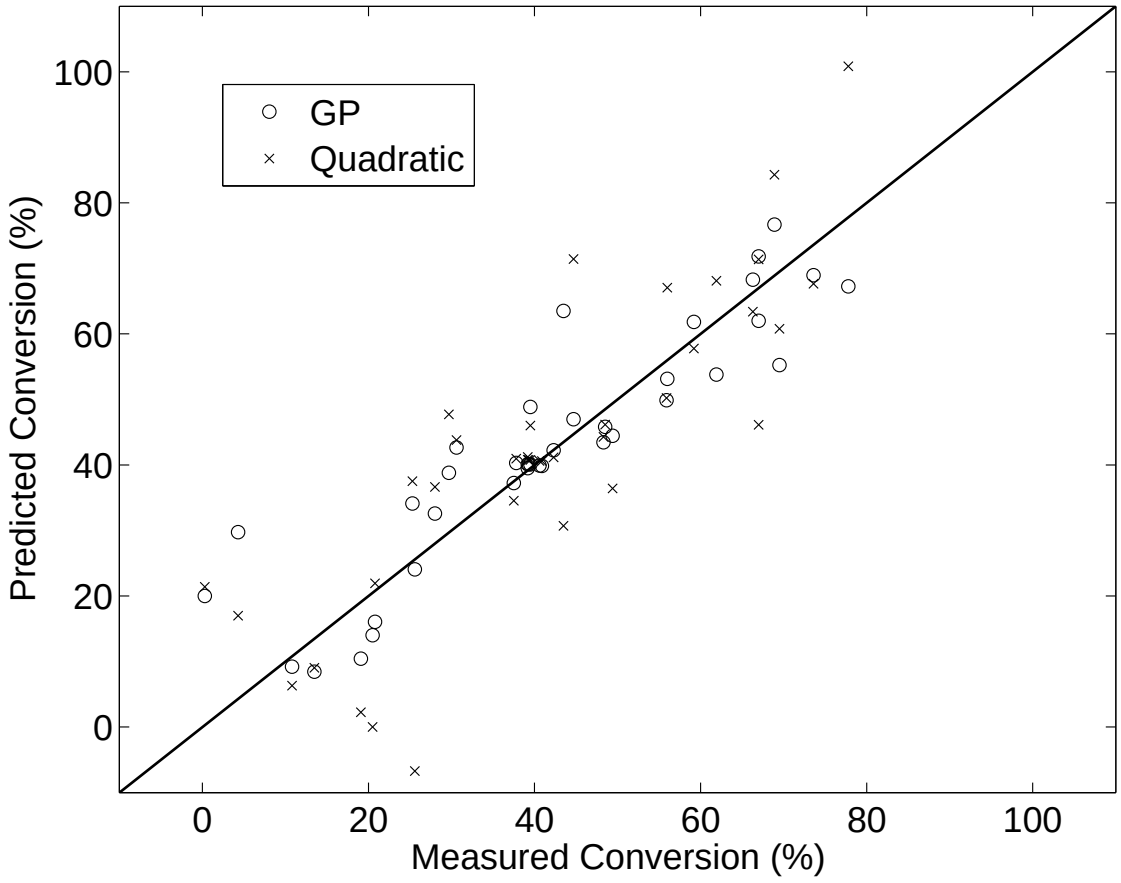

Figure 1: Leave-one-out cross-validation results for modeling benzyl alcohol conversion using GP ($\mathrm{RMSE_{cv}}$=8.36, R^2_{cv}=0.81) and quadratic regression ($\mathrm{RMSE_{cv}}$=12.35, R^2_{cv}=0.59) models.

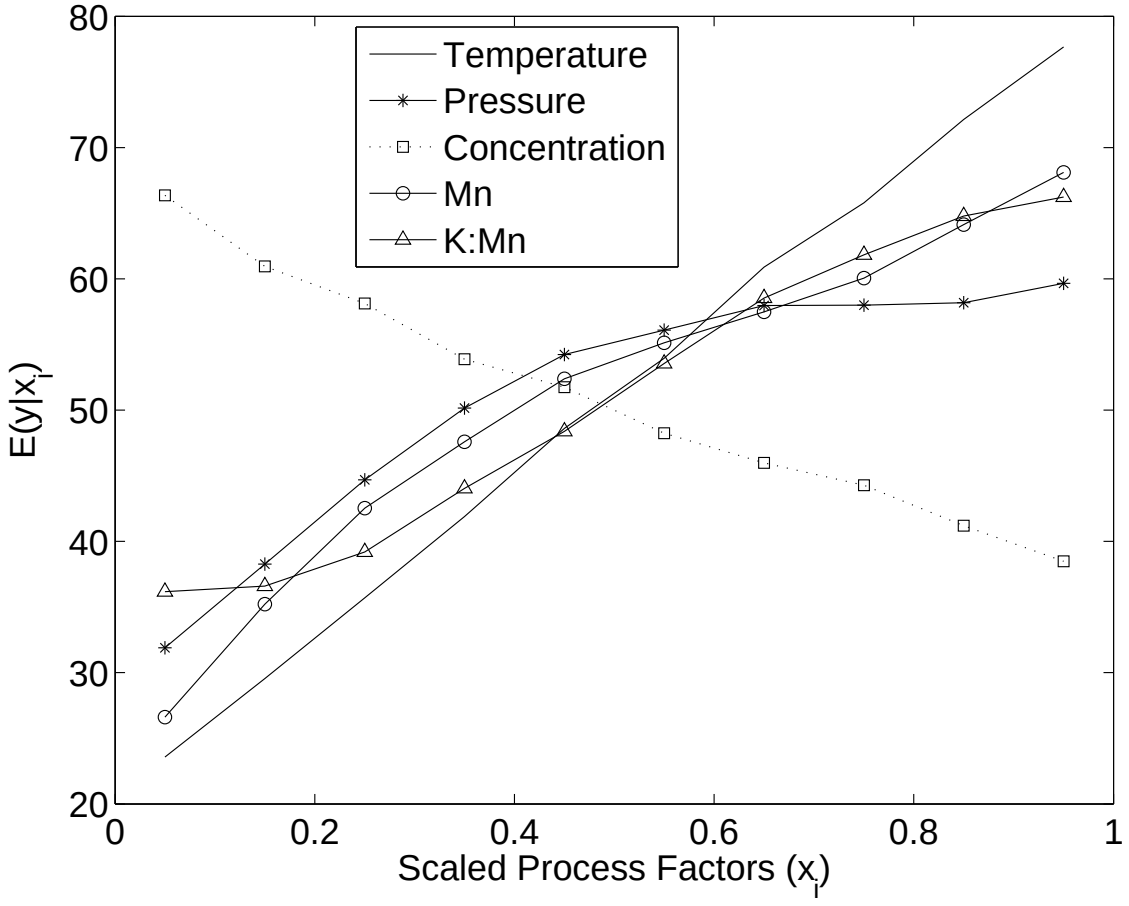

Figure 2: The main effect $E(y|x_i)$ against x_i for each process factor: the benzyl alcohol oxidation process.

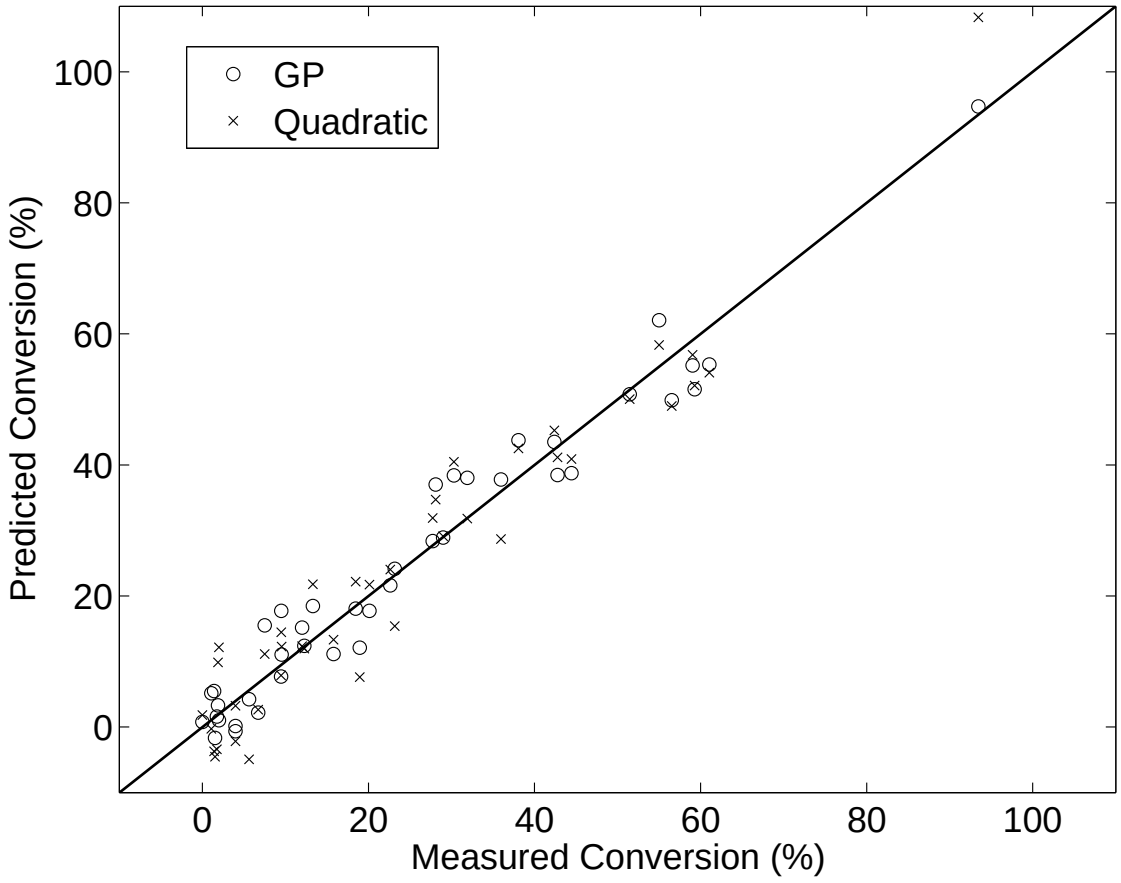

Figure 3: Leave-one-out cross-validation results for modeling *trans*-stilbene conversion using GP (RMSE_{cv}=4.50, R^2_{cv}=0.96) and quadratic regression (RMSE_{cv}=5.88, R^2_{cv}=0.92) models.

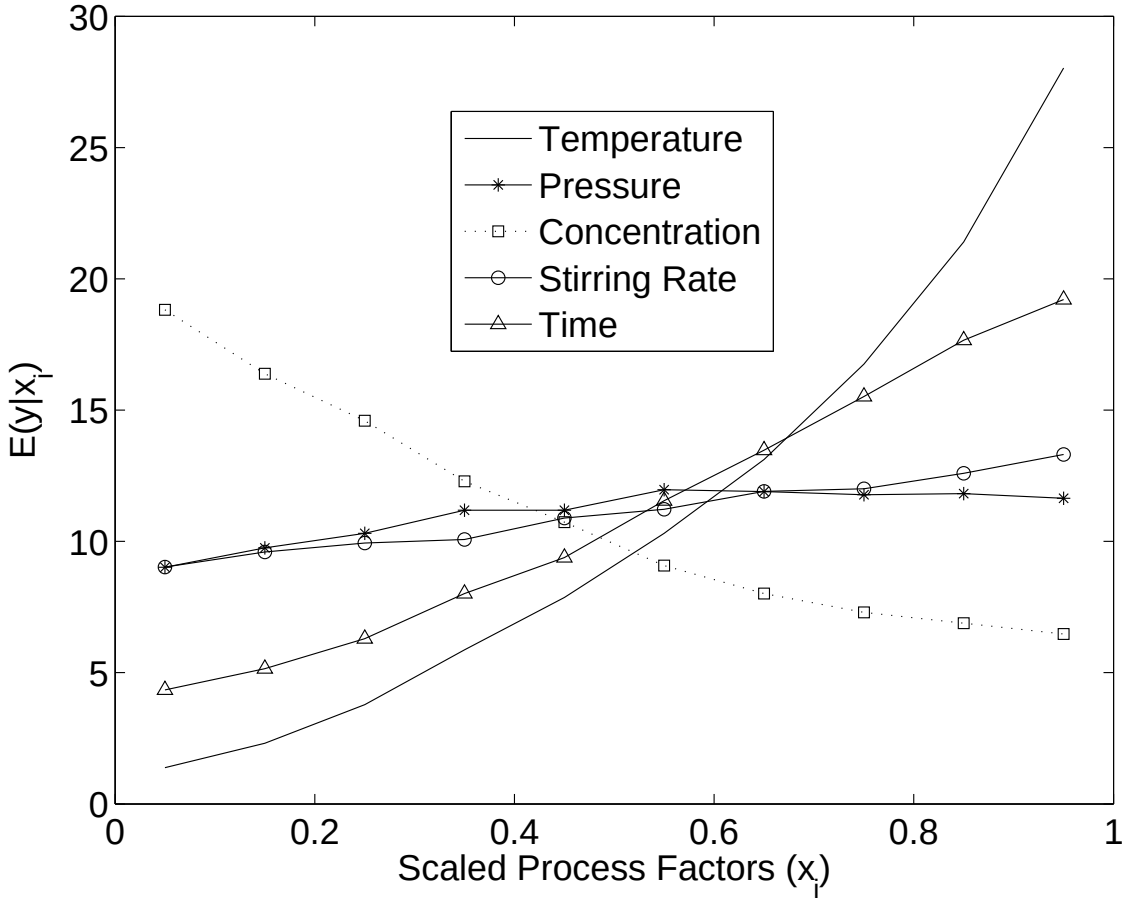

Figure 4: The main effect $E(y|x_i)$ against x_i for each process factor: the *trans*-stilbene epoxidation process.

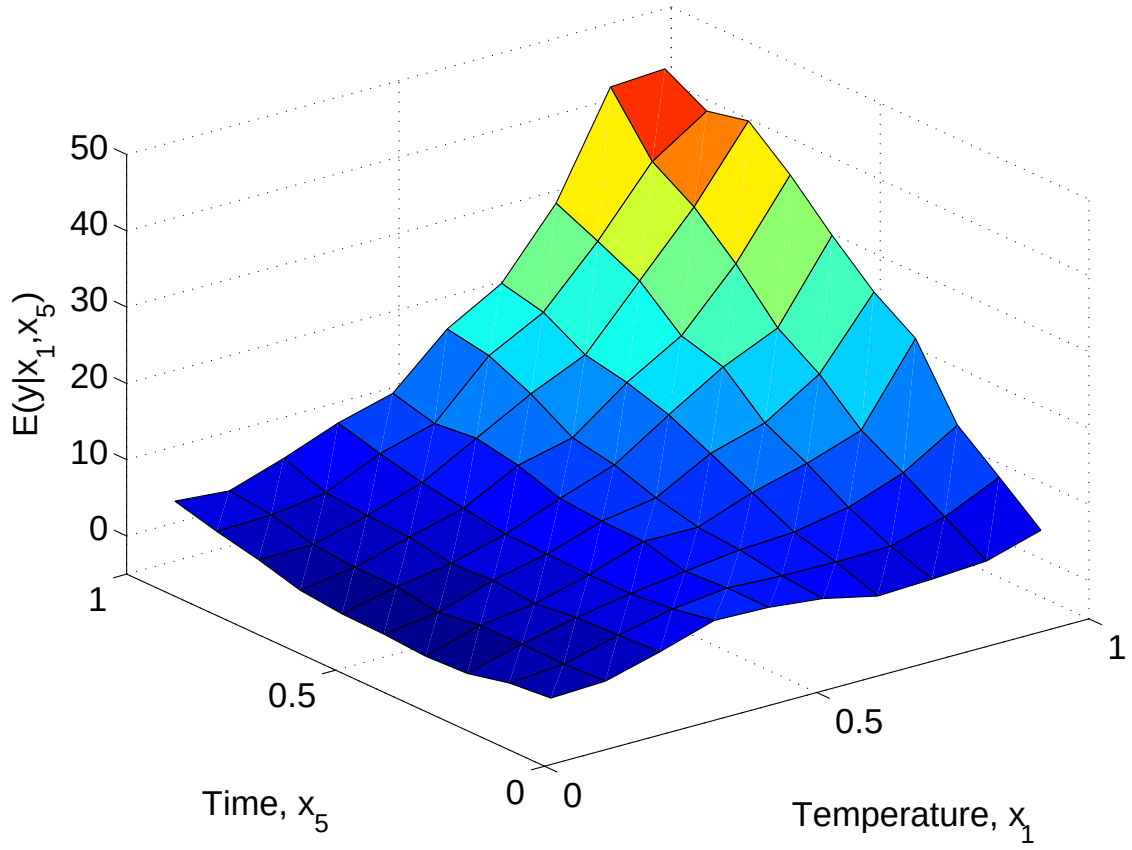

Figure 5: The expectation, $E(y|x_i, x_j)$, against reaction temperature (x_1) and reaction time (x_5) for the *trans*-stilbene epoxidation process. Process factors are scaled. The figure clearly shows positive interaction between the two process factors.

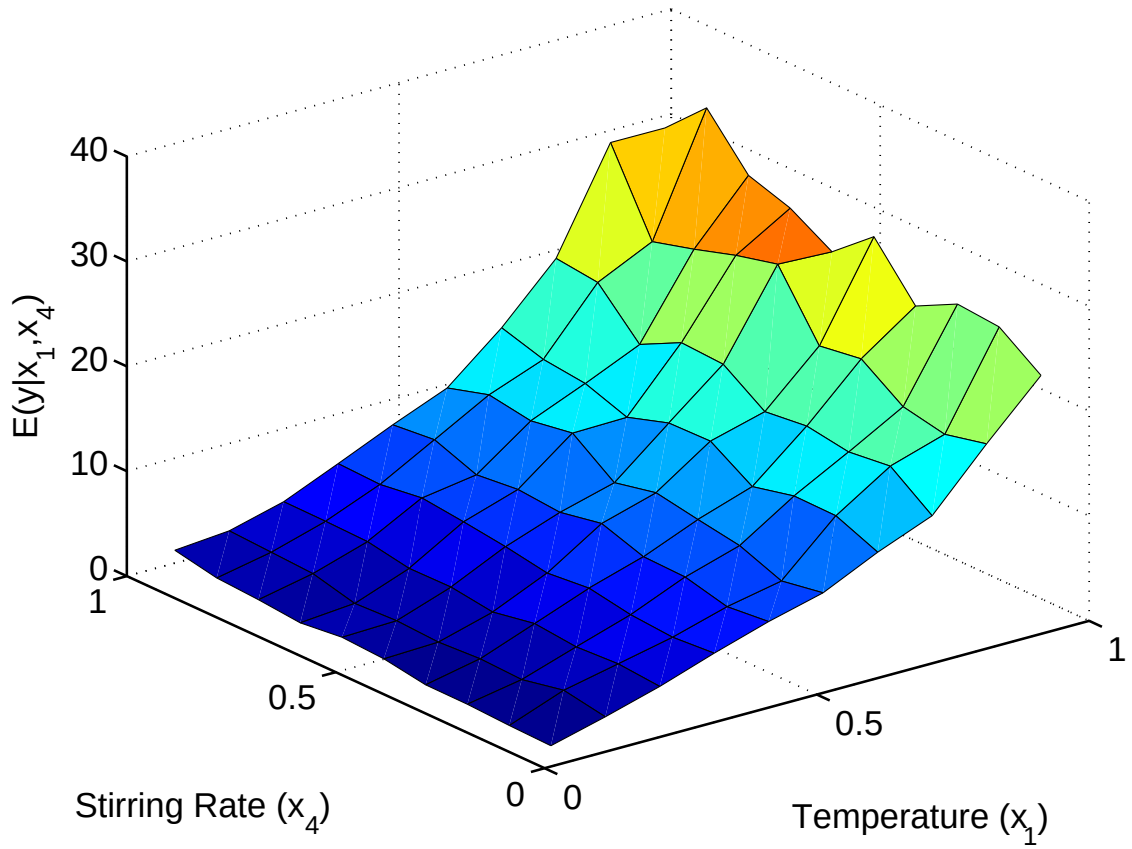

Figure 6: The expectation, $E(y|x_i, x_j)$, against reaction temperature (x_1) and stirring rate (x_4) for the *trans*-stilbene epoxidation process. Process factors are scaled. The figure does not suggest any interaction between the two process factors.