A Methodology for Combining Symbolic Regression and Design of Experiments to Improve Empirical Model Building

Flor Castillo, Kenric Marshall, James Green, and Arthur Kordon

The Dow Chemical Company
2301 N. Brazosport Blvd, B-1217
Freeport, TX 77541, USA
979-238-7554
{Facastillo,KAMarshall,JLGreen,Akordon}@dow.com

Abstract. A novel methodology for empirical model building using GP-generated symbolic regression in combination with statistical design of experiments as well as undesigned data is proposed. The main advantage of this methodology is the maximum data utilization when extrapolation is necessary. The methodology offers alternative non-linear models that can either linearize the response in the presence of Lack or Fit or challenge and confirm the results from the linear regression in a cost effective and time efficient fashion. The economic benefit is the reduced number of additional experiments in the presence of Lack of Fit.

1 Introduction

The key issues in empirical model development are high quality model interpolation and its extrapolation capability outside the known data range. Of special importance to industrial applications is the second property since the changing operating conditions are more a rule than an exception. Using linear regression models based on well-balanced data generated by Design of Experiments (DOE) is the dominant approach to effective empirical modeling and several techniques have been developed for this purpose [1]. However, in many cases, due to the non-linear nature of the system and unfeasible experimental conditions, it is not possible to develop a linear model. Among the several approaches that can be used either to linearize the problem, or to generate a non-linear empirical model is Genetic Programming (GP)-generated symbolic regression. This novel approach uses simulated evolution to generate non-linear equations with high fitness [2]. The potential of symbolic regression for linearizing the response in statistical DOE when significant Lack of Fit is detected and additional experimentation is unfeasible was explored in [3]. The derived transformations based on the GP equations resolved the problem of Lack of Fit and demonstrated good interpolation capability in an industrial case study in The Dow Chemical Company. However, the extrapolation capability of the derived non-linear models is unknown and not built-in as in the case of linear models.

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 1975–1985, 2003.

Extrapolation of empirical models is not always effective but often necessary in chemical processes because plants often operate outside the range of the original experimental data used to develop the empirical model. Of special importance is the use of this data since time and cost restrictions in planned experimentation are frequently encountered.

In this paper, a novel methodology integrating GP with designed and undesigned data is presented. It is based on a combination of linear regression models based on a DOE and non-linear GP-generated symbolic regression models considering interpolation and extrapolation capabilities. This approach has the potential to improve the effectiveness of empirical model building by maximizing the use of plant data and saving time and resources in situations where experimental runs are expensive or technically unfeasible. A case study with a chemical process was used to investigate the utility of this approach and to test the potential of model over-fitting. The promising results obtained give the initial confidence for empirical model development based on GP- generated symbolic regression in conjunction with designed data and open the door for numerous industrial applications.

2 A Methodology for Empirical Model Building Combining Linear and Non-linear Models

With the growing research in evolutionary algorithms and the speed, power and availability of modern computers, genetic programming offers a suitable possibility for real-world applications in industry. This approach based on GP offers four unique elements to empirical model building. First, previous modeling assumptions, such as independent inputs and an established error structure with constant variance, are not required [4]. Second, it generates a multiplicity of non-linear equations which have the potential of restoring Lack of Fit in linear regression models by suggesting variable transforms that can be used to linearize the response [3]. Third, it generates non-linear equations with high fitness representing additional empirical models, which can be considered in conjunction with linearized regression models or to challenge and confirm results even when a linear model is not significant [3]. Fourth, unlike neural networks that require significant amount of data for training and testing, it can generate empirical models with small data sets.

One of the most significant challenges of Genetic programming for empirical model building and linearizing regression models is that it produces non-parsimonious solutions with chunks of inactive terms (introns) that do not contribute to the overall fitness [4]. This drawback can be partially overcome by computational algorithms that quickly select the most highly fit and less-complex models or by using parsimony pressure during the evolution process.

Using DOE in conjunction with genetic programming results in a powerful approach that improves empirical model building and that may have significant economic impact by reducing the number of experiments.

The following are the main components of the proposed methodology.

Step 1. The Experimental Design

A well-planned and executed experimental design (DOE) is essential to the development of empirical models to understand causality in the relationship between the input

variables and the response variable. This component includes the selection of input variables and their ranges, as well as the response variable. It also includes the analysis and the development of a linear regression model for the response(s).

Step 2. Linearization via Genetic Programming
This step is necessary if the linear regression model developed previously presents Lack of Fit and additional experimentation to augment the experimental design such as a Central Composite Design (CCD) is not possible due to extreme experimental conditions. An alternative is to construct a Face-Centered Design by modification of the CCD [5]. However, this alternative often results in high correlation between the square terms of the resulting regression models. GP-generated symbolic regression can be employed to search for mathematical expressions that fit the given data set generated by the experimental design. This approach results in several analytical equations that offer a rich set of possible input transformations which can remove lack of fit without additional experimentation. In addition, it offers non-linear empirical models that are considered with the linearized regression model. The selection between these models is often a trade-off between model complexity and fitness. Very often the less complex model is easier to interpret and is preferred by plant personnel and process engineers.

Step 3. Interpolation
This step is a form of model validation in which the linearized regression model and the non-linear model(s) generated by GP are tested with additional data within the experimental region. Here again, model selection considers complexity, fitness, and the preference of the final user.

Step 4. Extrapolation
This is necessary when the range of one or more of the variables is extended beyond the range of the original design and it is desirable to make timely predictions on this range.

Step 5. Linear and Non-linear Models for Undesigned Data and Model Comparison
When the models developed in the previous sections do not perform well for extrapolation, it is necessary to develop a new empirical model for the new operating region. The ideal approach is to consider an additional experimental design in the new range but often it can not be done in a timely fashion. Furthermore, it is often desirable to use the data set already available. In these cases, while a multiple regression approach can be applied to build a linear regression model with all available data, the risk of collinearity, near linear dependency among regression variables, must be evaluated since the data no longer conforms to an experimental design. High collinearity produces ambiguous regression results making it impossible to estimate the unique effects of individual variables in the regression model. In this case regression coefficients have large sampling errors and are very sensitive to slight changes in the data and to the addition or deletion of variables in the regression equation. This affects model stability, inference and forecasting that is made based on the regression model. Of special interest here is a model developed with a GP-generated symbolic regression algorithm because it offers additional model alternatives.

One essential consideration in this case is that the models generated with this new data *can not* be used to infer *causality*. This restriction comes from the fact that the

data no longer conforms to an experimental design. However, the models generated (linear regression and non-linear GP-generated models) *can* be used to *predict* the response in the new range. Both types of models are then compared in terms of complexity, and fitness, allowing the final users to make the decision.

The proposed methodology will be illustrated with an industrial application in a chemical reactor.

3 Empirical Modeling Methodology for a Chemical Reactor

3.1 The Experimental Design and Transformed Linear Model (Steps 1 and 2 in Methodology)

The original data set corresponds to a series of designed experiments that were conducted to clarify the impact on formation of a chemical compound as key variables are manipulated. The experiments consisted of a complete 2^4 factorial design in the factors x_1, x_2, x_3, x_4 with three center points. Nineteen experiments were performed. The response variable, S_k, was the yield or selectivity of one of the products. The factors were coded to a value of −1 at the low level, +1 at the high level, and 0 at the center point. The complete design in the coded variables is shown in Table 1, based on the original design in [3].

The selectivity of the chemical compound (S_k), was first fit to the following first-order linear regression equation considering only terms that are significant at the 95% confidence level:

$$S_k = \beta_o + \sum_{i=1}^{k} \beta_i x_i + \sum \sum_{i<j} \beta_{ij} x_i x_j \tag{1}$$

Table 1. 2^4 factorial design with three center points

Runs	x_1	x_2	x_3	x_4	S_k	P_k
1	1	-1	1	1	1.598	0.000
2	0	0	0	0	1.419	0.000
3	0	0	0	0	1.433	0.016
4	-1	1	1	1	1.281	0.016
5	-1	1	-1	1	1.147	0.009
6	1	1	-1	1	1.607	0.012
7	-1	1	1	-1	1.195	0.019
8	1	1	1	-1	2.027	0.015
9	-1	-1	-1	1	1.111	0.009
10	-1	1	-1	-1	1.159	0.007
11	-1	-1	-1	-1	1.186	0.006
12	1	-1	-1	1	1.453	0.013
13	1	1	-1	-1	1.772	0.006
14	-1	-1	1	-1	1.047	0.018
15	-1	-1	1	1	1.175	0.009
16	1	1	1	1	1.923	0.023
17	1	-1	-1	-1	1.595	0.007
18	1	-1	1	-1	1.811	0.015
19	0	0	0	0	1.412	0.017

Lack of Fit was induced (p = 0.046) by omission of experiment number 1 of the experimental design. This was done to simulate a common situation in industry in which LOF is significant and additional experimental runs are impractical due to the cost of experimentation, or because it is technically unfeasible due to restrictions in experimental conditions.

The GP algorithm (GP-generated symbolic regression) was applied to the same data set. The algorithm was implemented as a toolbox in MATLAB. Several models of the selectivity of the chemical compound as a function of the four experimental variables (x_1, x_2, x_3, x_4) were obtained by combining basic functions, inputs, and numerical constants. The initial functions for GP included: addition, subtraction, multiplication, division, square, change sign, square root, natural logarithm, exponential, and power. Function generation takes 20 runs with 500 population size, 100 number of generations, 4 reproductions per generation, 0.6 probability for function as next node, 0.01 parsimony pressure, and correlation coefficient as optimization criteria.

The functional form of the equations produced a rich set of possible transforms. The suggested transforms were tested for the ability to linearize the response without altering the necessary conditions of the error structure needed for least-square estimation (uncorrelated and normally distributed errors with mean zero, and constant variance). The process consisted of selecting equations from the simulated evolution with correlation coefficients larger than 0.9. These equations were analyzed in terms of the R^2 between model prediction and empirical response. Equations with R^2 higher than 0.9 were chosen and the original variables were transformed according to the functional form of these equations. The linear regression model presented in equation (1) was fitted to the data using the transformed variables and the fitness of this transformed linear model was analyzed considering Lack of Fit and R^2. The error structure of the models not showing lack of fit was then analyzed. This process ensured that the transformations given by GP not only linearized the response but also produced the adequate error structure needed for least square estimations.

Using the process previously described, the best fit between model prediction and empirical response from the set of potential non-linear equations was found for the following analytical function with R^2 of 0.98[3]:

$$S_k = \frac{3.13868\times10^{-17} e^{\sqrt{2x_1}} \ln\left[(x_3)^2\right] x_2}{x_4} + 1.00545 \qquad (2)$$

The corresponding input/output sensitivity analysis reveals that x_1 is the most important input.

The following transformations were then applied to the data as indicated by the functional form of the GP function shown in equation (2).

Table 2. Variable transformations suggested by GP model.

Original Variable	Transformed Variable
x_1	$Z_1 = \exp\left(\sqrt{2x_1}\right)$
x_2	$Z_2 = x_2$
x_3	$Z_3 = \ln[(x_3)^2]$
x_4	$Z_4 = x_4^{-1}$

The transformed variables were used to fit a first-order linear regression model shown in equation (1). The resulting model is referred to as the Transformed Linear Model (TLM). The TLM had an R^2 of 0.99, no evidence of Lack of Fit (p=0.3072) and retained the appropiate randomized error structure indicating that the GP-generated transformations were succesful in linearizing the response. The model parameters in the transformed variables are given in [3].

3.2 Interpolation Capability of Transformed Linear Model (TLM) and Symbolic Regression model (GP) (Step 3 of Methodology)

The interpolation capabilities of the TLM and the GP model shown in equation (2) were tested with nine additional experimental points within the range of experimentation. A plot of the GP and TLM models for the interpolation data is presented in Fig. 1.

The models, GP and TLM, perform similarly in terms of interpolation with sum square errors (SSE) being slightly smaller for the TLM (0.1082) than for the GP model (0.1346). However, the models are comparable in terms of prediction with data within the region of the design. The selection of one of these models would be driven by the requirements of a particular application. For example, in the case of process control, the more parsimonious model would generally be preferred.

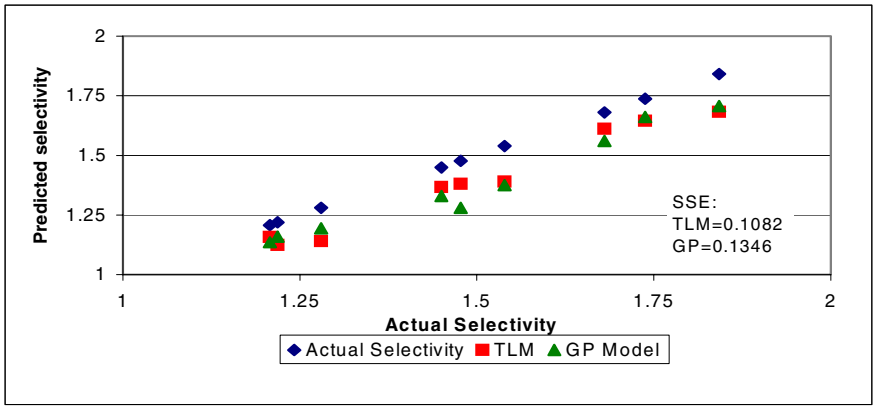

Fig. 1. Predicted selectivity for GP and TLM using interpolation data set

3.3 Extrapolation Capability of Transformed Linear Model (TLM) and Symbolic Regression Model (GP) (Step 4 of Methodology)

It was necessary to evaluate the prediction of the TLM and the GP model with available data outside the region of experimentation (beyond –1 and 1 for the input variables) that were occasionally encountered in the chemical system. The most appropriate approach would be to generate an experimental design in this new region of operability. Unfortunately this could not be completed in a timely and cost effective

fashion due to restrictions in operating conditions. The data used for extrapolation in coded form is shown in Table 3. Figure 2 shows the comparison between actual and predicted selectivity for this data set.

Table 3. Data used for extrapolation for GP and TLM models

Run	X1	X2	X3	X4	Selectivity S.	GP	TLM	SSE GP	SSE TLM
1	1.0	-	-	-	1.614	1.63	1.61	0.000	0.000
2	1.0	-	-	-	1.331	1.32	1.21	0.000	0.013
3	1.0	-	-	-	1.368	1.36	1.40	0.000	0.001
4	2.0	-	-	-	1.791	2.27	2.02	0.231	0.056
5	2.0	-	-	-	1.359	1.65	1.24	0.086	0.013
6	2.0	-	-	-	1.422	1.72	1.47	0.092	0.003
7	3.0	-	-	-	1.969	3.52	2.63	2.412	0.446
8	3.0	-	-	-	1.398	2.29	1.28	0.798	0.013
9	3.0	-	-	-	1.455	2.43	1.57	0.960	0.014
10	3.0	0.67	-	-	1.480	2.62	2.10	1.318	0.384
11	3.0	0.76	-	0.47	1.343	2.30	1.48	0.923	0.020
Data is coded based on conditions of the original design								6.822	0.963

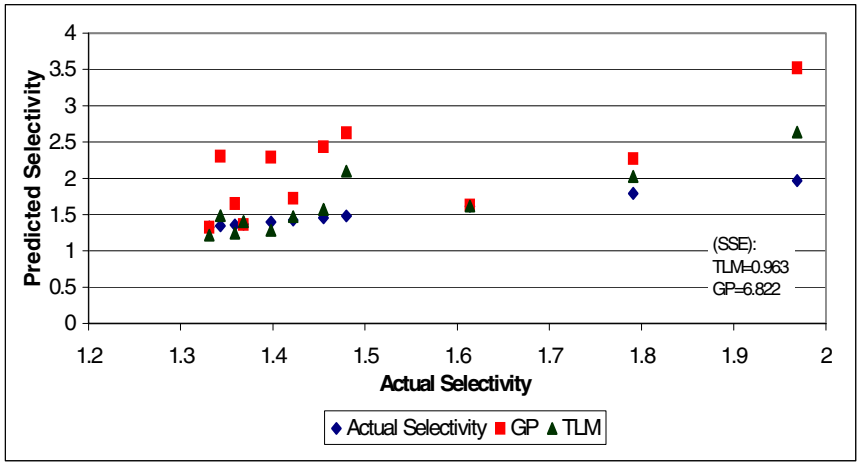

Fig. 2. Predicted selectivity for GP and TLM models using extrapolation data

The deviations between model predictions and actual selectivity (SSE) are larger for the GP model (6.822) than for the TLM (0.963) suggesting that the TLM had better extrapolation capabilities. Table 3 shows that the sum squares errors (SSE) gets larger for both models as the most sensitive input x_1 increases beyond 1,which is the region of the original design. This finding confirms that extrapolation of a GP model in terms of the most sensitive inputs is a dangerous practice. This is not a disadvantage of the GP or TLM per se, it is a fact that has challenged and will always challenge empirical models, whether linear or non-linear in nature. This is because, unlike mechanistic models that are *deduced* from the laws of physics and which apply in

in general for the phenomena being modeled, empirical models are *developed* from data in a limited region of experimentation.

3.4 Linear Regression and Symbolic Regression for Undesigned Data (Step 5 of Methodology)

Given that extrapolation of the previously developed models was not effective, a different approach was explored by combining all of the data sets previously presented (DOE, interpolation and extrapolation data sets) to build a linear regression model and a GP model. However since the combined data sets do not conform to an experimental DOE, the resulting models are only to be used for prediction within the new region of operability and not to infer causality.

The Linear Regression Model. The linear regression model was built treating the three sets as one combined data set and treating them as undesigned data, or data not collected using a design of experiments. Selectivity was fit to the first-order linear regression shown in equation (1). The resulting model is referred to as Linear Regression Model (LRM). The analysis of variance revealed a significant regression equation (F ratio <0.0001) with R^2 of 0.94, and no evidence of Lack of Fit (p = 0.1520). The parameter estimates are presented in Table 4.

Table 4. Parameter Estimates for Linear Regression Model showing significant terms at 95%

Term	β Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-	0.365	-	<.000	
x1	0.008585	0.000	17.49	<.000	1.871
x2	0.086230	0.033	2.61	0.014	1.506
x3	0.105926	0.010	9.72	<.000	1.995
x4	-	0.119	-4.21	0.000	1.688
(x1-615.789)*(x2-3.73)	0.002362	0.001	2.17	0.038	1.748
(x1-615.789)*(x3-4.37447)	0.003111	0.000	10.63	<.000	1.537
(x2-3.73)*(x3-4.37447)	0.045254	0.023	1.94	0.062	1.791
(x1-615.789)*(x4-1.18474)	-	0.003	-2.47	0.020	1.695
(x3-4.37447)*(x4-1.18474)	0.036342	0.074	0.49	0.628	1.605
(x1-615.789)*(x3-4.37447)*(x4-	-	0.001	-2.80	0.009	1.766
effect included to preserve model precedence given that third order interaction is significant					

Because the LRM was built for undesigned data, multicollinearity was analyzed using Variance Inflation Factors (VIF) [6]. As previously discussed, the presence of severe multicollinearty can seriously affect the precision of the estimated regression coefficients, making them very sensitive to the data in the particular sample collected and producing models with poor prediction.

The VIF for each model parameter (VIF_j) were compared to the overall Model (VIF_{model}). The VIF for the LRM are listed in Table 4. No evidence of severe multicollinearity was detected (VIF_j < 16.7526). A subsequent residual analysis did not

show indication of violations of the error structure required for least square estimation indicating that an acceptable LRM was achieved with the combined data set.

The GP Model. The GP model was developed for the combined data set by using a toolbox in MATLAB. This model is referred to as the GP2 model to differentiate it from the GP model that was developed using only the DOE data. The initial functions for GP2 included addition, subtraction, multiplication, division, square, change sign, square root, natural logarithm, exponential, and power. Several runs were performed with various function generation (20 to 40); population size (500 to 900); number of generations (100 to 500); parsimony pressure (0.01 to 0.1); and 40 reproductions per generation with 0.6 probability for function as next node, and two different optimization criteria (correlation coefficient and sum of squares).

Several hundred equations were obtained. The selection of the best equations was based on the value of R^2. The following equation resulted in the highest value of R^2 (0.84).

$$S_k = -4.5333 - 0.008\,\mathrm{I}*\left(-e^{2x_4\left[e^{-2x_3*(-x_2+x_3)}\right]}*\left|1.9997\,\mathrm{H}+\sqrt{\left|\sqrt{|x_1|}\right|^2}-\sqrt{|x_1|}+e^{2\mathrm{Re}\left[\frac{x_3+x_4}{\log\left(e^{x_3}\right)}\right]}\right.\right.*\left.\left.|x_4|^2-x_2(5.052\,1\mathrm{H}+x_3-x_4)\right.\right) \quad (3)$$

Unlike the model of equation (1), the GP2 model shown in equation (3) is a non-linear model, with a complex functional form that shows relationships between the different variables (x_1, x_2, x_3, x_4).

Comparing the Linear and Non-linear Empirical Models. Figure 3 shows the graph of the GP2 model and the Linear Regression Model (LRM) for the combined data set. Comparing the two empirical models, the LRM performs better that the GP2 model in terms of R^2 and the sum square errors.

In terms of influential observations, the linear regression model allows determination of influential observations by calculating the Cook's D influence, D_i [7]. Observations with large values of D_i have considerable influence on the least square estimates β_i in equation (1) making these estimates very sensitive to these observations. None of the observations in the combined data set was considered influential (all calculated $D_i<1$). This information is helpful identifying potentially interesting observations that may be replicated to confirm results.

The GP algorithm does not allow the statistical determination of influential observations but it allows the determination of the most sensitive inputs. This process is somehow similar to the testing of significant parameters in linear regression but it is not based on *statistical hypothesis testing*. The GP algorithm finds the most sensitive inputs by determining how the fitness function of the GP-generated equations improves by adding the specific input. The input/output sensitivity analysis had shown x_1 as the most important input. This is in agreement with the LRM, which shows x_1 as significant (Table 4).

The simpler form of the LRM would generally be preferred by plant engineers because of the larger R^2 and because it is more parsimonious than the GP2 model shown

in equation (3). Nevertheless, the functional form of the non-linear GP2 model reveals relationships among the variables that account for a large percentage of the variation of the data.

Therefore, designed and undesigned data are opportunities for consideration of an alternative GP generated model. Even when the regression model is not significant, a GP model can still be built to confirm or challenge the result. This is illustrated in the following section.

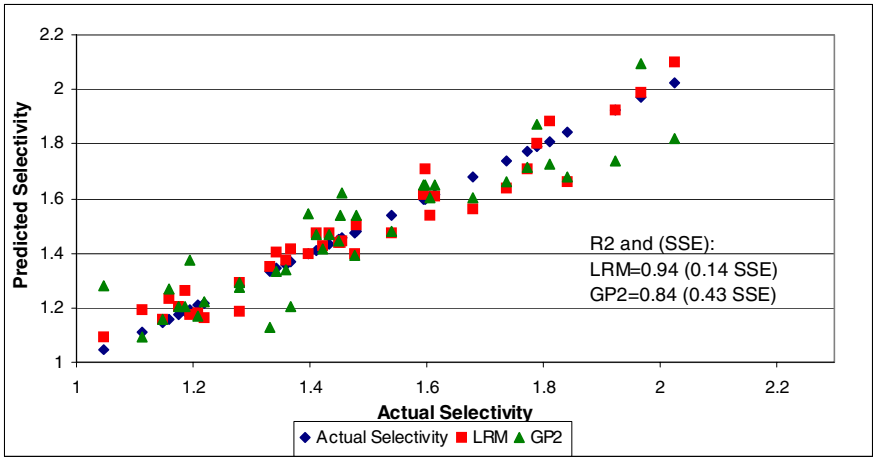

Fig. 3. Predicted selectivity for GP2 and LRM models for combined data set data

3.5 The GP Model – LRM Not Significant

A known weakness of some non-linear algorithms is the risk of model over-fitting (modeling unexplainable variation instead of real relationships that exist between inputs and outputs). This risk can significantly hinder the potential of GP models.

In order to test model over-fitting, data for the selectivity of a second chemical compound (P_k) was selected from the same 2^4 DOE experiments shown in Table 1 (P_k selectivity data is shown in the fourth column in Table 1). The standard linear regression model shown in equation (1) was used to fit to the data considering selectivity P_k as the response variable; and x_1, x_2, x_3, x_4 as the independent variables.

The corresponding analysis of variance indicated no evidence of Lack of Fit at the 95% confidence (p=0.0938) but revealed a non-significant model (p=0.4932) indicating that the hypothesis that all regression coefficients βi in the model of equation (1) are zero could not be rejected. Thus, the linear model is reduced to the mean.

The GP algorithm has been applied to the DOE data set. The selectivity P_k was selected as the output variable and x_1, x_2, x_3, x_4 were selected as input variables. Several runs were performed with 20 runs, population size between 500 to 900, number of generations between 100 to 500, 40 reproductions per generation, 0.6 probability for function as next node, parsimony pressure between 0.01 and 0.1, and correlation coefficient and sum of squares as optimization criteria. Hundreds of non-linear GP equa-

tions were generated. However, no equation was found that accounted for more than 50% of the variation in the data (the maximum correlation coefficient found was 0.5).

This result is potentially significant. In the P_k case, it can be demonstrated through statistical analysis that a statistically significant correlation does not exist between variables and response. The non-linear GP algorithm produced in an independent way the same result from the linear regression analysis by not creating a statistically significant model. This demonstrates that in this case the risk of model over-fitting is low.

4 Conclusions

A novel methodology for empirical modeling based on Design of Experiments and GP has been defined and applied successfully in the Dow Chemical Company. The proposed methodology uses linear regression models and non-linear GP models (statistically designed experiments, linear regression models for undesigned data, and GP-generated symbolic regression). A significant part of this methodology is based on the unique potential of GP algorithms for linearizing regression models in the presence of Lack of Fit and providing additional empirical non-linear functions that can be considered in combination with the linear ones. The proposed methodology has the following advantages:

- increases the options for development of empirical models based on DOE;
- reduces (or even eliminates) the number of additional experiments in the presence of Lack of Fit;
- maximizes the use of available data when model extrapolation is required;
- improves model validation by introducing alternative models.

These advantages are illustrated in an industrial application. The promising results obtained constitute a solid foundation for utilization of linear and non-linear models in industrial applications where extrapolation of an empirical model is often required.

References

1. Box, G.E.P, and Draper, N. R., *Empirical Model Building and Response Surfaces*, John Wiley and Sons, New York, 1987.
2. Koza, J. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA, 1992.
3. Castillo F. A.,Marshall, K. A, Green, J.L., and Kordon, A, "Symbolic Regression in Design of Experiments: A Case Study with Linearizing Transformations, *Proceedings of GECCO'2002*, New York, pp. 1043–1048., 2002
4. Banzhaf W., P. Nordin, R. Keller, and F. Francone, *Genetic Programming: An Introduction*, Morgan Kaufmann, San Francisco, 1998.
5. Myers, R.H., and, Montgomery, D.C., *Response Surface Methodology*, John Wiley and Sons, New York, 1995
6. Montgomery, D.C., and Peck, E.A., *Introduction to Linear Regression Analysis*, John Wiley and Sons, New York, 1992.
7. Cook, R.D., "Detection of Influential Observations in Linear Regression", *Technometrics*, Vol. 19, pp. 15–18, 1977