# Pareto Front Genetic Programming Parameter Selection Based on Design of Experiments and Industrial Data

Flor Castillo
The Dow Chemical Company
2301 N Brazosport Blvd, B1217
Freeport, TX 77541
facastillo@dow.com

Arthur Kordon
The Dow Chemical Company
2301 N Brazosport Blvd, B1217
Freeport, TX 77541
akordon@dow.com

Guido Smits
Dow Benelux, B.V.
Herbert H. Dowweg 5, POB 48
Terneuzen, The Netherlands, 4530 AA
gfsmits@dow.com

Ben Christenson
The Dow Chemical Company
2301 N Brazosport Blvd, B2018
Freeport, TX 77541
bchristenson@dow.com

Dee Dickerson
The Dow Chemical Company
2301 N Brazosport Blvd, B2018
Freeport, TX 77541
ddickerson2@dow.com

**ABSTRACT**

Symbolic regression based on Pareto Front GP is the key approach for generating high-performance parsimonious empirical models acceptable for industrial applications. The paper addresses the issue of finding the optimal parameter settings of Pareto Front GP which direct the simulated evolution toward simple models with acceptable prediction error. A generic methodology based on statistical design of experiments is proposed. It includes statistical determination of the number of replicates by half-width confidence intervals, determination of the significant inputs by fractional factorial design of experiments, approaching the optimum by steepest ascent/descent, and local exploration around the optimum by Box Behnken or by central composite design of experiments. The results from implementing the proposed methodology to a small-sized industrial data set show that the statistically significant factors for symbolic regression, based on Pareto Front GP, are the number of cascades, the number of generations, and the population size. A second order regression model with high $R^2$ of 0.97 includes the three parameters and their optimal values have been defined. The optimal parameter settings were validated with a separate small sized industrial data set. The optimal settings are recommended for symbolic regression applications using data sets with up to 5 inputs and up to 50 data points.

## Categories and Subject Descriptors

G.3. [**Mathematics of Computing**]: Probability and statistics–Correlation and regression analysis.

I.2.2 [**Artificial Intelligence**]: Automatic programming and program synthesis

**General Terms:** Experimentation. Performance

**Keywords:** Genetic Programming, Pareto front, statistical design of experiments, symbolic regression, industrial applications



## 1. INTRODUCTION

One of the issues any researcher and practitioner needs to resolve dealing with Genetic Programming (GP) is to select a proper set of parameters, such as number of generations, population size, crossover probability, mutation rate, etc. Surprisingly, there are very few investigations on this topic known in the literature. In his first book, Koza [1] gives several rules of thumb for parameter selection based on simulation experience. Similar recommendations are given in Banzhaf et al. [2]. The only statistically-based study by Felt and Nordin [3] investigated the effect of 17 GP parameters on three binary classification problems using highly fractionated experimental statistical designs assuming, in some cases, that even second- and third-order interactions are not significant, i.e., the combined effect of two factors and three factors has no effect on the response. However, these assumptions have not been verified. Recently, Petrovski *et al* [4] investigated the performance of genetic algorithms using experimental design and optimization techniques.

However, the growing interest of industry in GP [5] requires a more systematic approach for the GP model generation process to guarantee consistency of delivered results. An important part of this process is the appropriate parameters setting for each specific type of applications, which will improve the efficiency of model development and minimize the development cost. The best way to address this issue is by using statistical Design Of Experiments (DOE) [6] on industrial data. Fortunately, as a result of the current successful GP applications, a set of industrial benchmark data sets has been collected (a summary for the applications in the chemical industry is given in [5]). They are with different sizes and data quality and each one is

a source of a successful real world application based on GP-generated symbolic regression.

The paper describes the statistical methodology and the results for finding the optimal parameter settings of a specific type of GP, called Pareto-Front GP, based on multi-objective optimization [7]. The results are within the scope of symbolic regression applications. The paper is organized in the following manner. First, the specific features of the Pareto Front GP approach, its importance to industrial applications, and the selected setting parameters are discussed in section 2. The DOE methodology for Pareto-Front GP parameter selection, which in addition to the experimental design includes the assessment of the necessary statistically significant repeats, is described in Section 3. The results for finding optimal GP parameter settings for a small-size industrial data set are given in Section 4 with validation on a different small-scale industrial data set, shown in Section 5.

## 2. KEY PARAMETERS OF PARETO FRONT GENETIC PROGRAMMING FOR SYMBOLIC REGRESSION

One of the areas where GP has a clear competitive advantage in real world applications is fast development of nonlinear empirical models [5]. However, if the GP-generated functions are based on high accuracy only, the high-fitness models are very complex, difficult to interpret, and crash in even minor changes in operating conditions. Manual selection of models with lower complexity and acceptable accuracy requires time consuming screening through large number of models. The solution is by using multi-objective optimization to direct the simulated evolution toward simple models with sufficient accuracy. Recently, a special version of GP, called Pareto-front GP, has significantly improved the efficiency of symbolic-regression model development, which has been demonstrated in several industrial applications [7]. In Pareto-front GP the simulated evolution is based on two criteria – prediction error (for example, based on $1-R^2$) and complexity (for example, based on the number of nodes). The optimal models fall on the curve of the non-dominated solutions, called Pareto front, i.e., no other solution is better than the solutions on the Pareto front in both complexity and performance. Of special importance to industry are the most parsimonious models with high performance, which occupy the lower left corner of the Pareto front (see Fig. 1). From that perspective, the objective of parameter settings is to select GP parameters that push the simulated evolution toward the parsimonious models with high performance, i.e. to guarantee a consistent convergence to the lower left corner of the Pareto front.

For the statistical DOE we need to define the target or response variable and the independent parameters or factors we would like to explore. The response variable proposed is the percentage of the area below the Pareto front (see Fig.1). In this case the accuracy or the prediction error is calculated as $(1-R^2)$ and the complexity is represented by the sum of the number of nodes of all sub-equations [7]. From practical consideration, an upper complexity limit of 400 is defined. The Pareto front line is obtained by interpolating through the points on the Pareto front. The area below the Pareto front is calculated within the limits of prediction error between 0 and 1 and complexity between 0 and 400. This area is divided on the full rectangle area and the response is the calculated percentage. For example, the response of Pareto Front 1 is 23% and the response of Pareto Front 2 is 56% (see Fig.1). The objective of the DOE is to select factors that minimize the response, i.e., to push the Pareto Front toward the origin where the simple models with low prediction error are located.

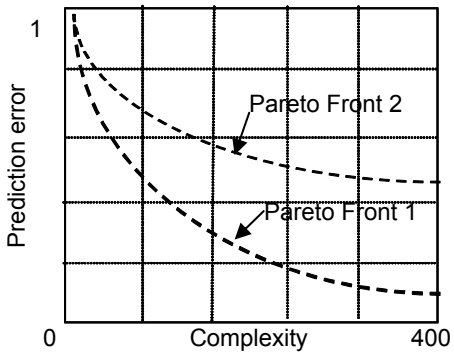


**Figure 1.** Responses for Pareto Front GP

The selected Pareto front GP parameters (factors) and their ranges are presented in the following table:

**Table 1. Factors for the Pareto Front GP DOE**

| Factor | Low level (-1) | High Level (+1) |
|---|---|---|
| $x_1$ - Number of cascades | 10 | 50 |
| $x_2$ - Number of generations | 10 | 50 |
| $x_3$ - Population size | 100 | 500 |
| $x_4$ - Probability of function selection | 0.4 | 0.65 |
| $x_5$ - Size of archive in % of pop. size | 50 | 100 |

One of the key features of Pareto Front GP is the availability of an archive to save the Pareto Front models during the simulated evolution. This creates two measures for the duration of simulations. The first measure is based on starting conditions of randomization of the population and the archive and the duration of the whole simulation is called a run. The second measure, called a cascade, is based on randomization of the population but the content of the archive is kept and participates in the evolution, i.e. it's possible to have several cascades in a single run. The number of cascades is the first factor in the DOE and it reflects the number of independent runs with a freshly generated starting population and kept the Pareto Front models in the archive. Factor $x_2$ represents the number of generations, factor $x_3$ is the population size, factor $x_4$ determines the probability of function selection and factor $x_5$ defines the archive size in percentage of the population size. The ranges of the factors have been selected based on the experience from various types of practical problems, related to symbolic regression. Since the objective is a consistent

Pareto front close to the origin, they differ from the recommendations for the original GP.

## 3. STATISTICAL DESIGN OF EXPERIMENTS FOR PARETO FRONT GP

Design of Experiments is a statistical approach that allows to further enhancing the knowledge of a system by quantifying the effect of a set of inputs (factors) on an output (response). This is accomplished by systematically running experiments at different combinations of the factor settings [6].

A classical DOE is the $2^k$ design in which all factors are investigated at an upper and lower level of a range resulting in $2^k$ experiments where *k* is the number of factors. This design has the advantage that the effects of the individual factors (main effects) as well as all possible interactions (combination of factors) can be estimated. However the number of experimental runs increases rapidly as the number of factors increases. If the number of experiments is impractical, fractional factorial ( FF) design can be used. In this case only a fraction of the full $2^k$ design is run by assuming that some interactions among factors are not significant. However in this case the main effects and interactions are confounded (cannot be estimated separately).

Depending on the type of fractional factorial, main effects may be confounded with second-, third-, or fourth-order interactions. The level of confounding is dictated by the design resolution. The higher the design resolution, the less confounding among factors. For example, a resolution III design confounds main effects with second-order interactions; a resolution IV design confounds second-order interaction with other second-order interactions; and a resolution V design confounds second-order interactions with third-order interactions.

When the objective of experimentation is to find the values of the inputs that will yield a maximum or a minimum for a specific response, the DOE strategy used is known as *response surface*. In this case fractional factorials are initially used to determine if the initial setting of the inputs are far from the desired optimum. This initial design is used to determine new levels of the inputs which approach to the optimum (this is known as the steepest ascent/descent path) [11]. As the optimum is approached other DOE techniques such as central composite designs (CCD) are used for local explorations so that the optimum can be identified and the conditions for practical used can be determined [6].

Applying DOE techniques to determine the optimum set of GP parameters needs to address also the issue of the statistically significant number of replications (independent runs). This is of key importance because the variability of the response may not be the same for the different combination of factors. To estimate the number of required replications, the half width (HW) confidence interval method [8] can be used.

The half width (HW) is defined as:

$$HW = t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \quad , \qquad (1)$$

where $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ percentage point of the *t* distribution with *n-1* degrees of freedom, *S* is the standard deviation and *n* is the number of runs.

A plot of the 100(1-α)% HW confidence interval reveals the minimum number of replications for a determined value of HW. 100(1-α)% *confidence interval* is a range of values in which the true answer is believed to lie with 1- α probability. Usually α is set at 0.05 so that 95% confidence interval is calculated. Half width, sometimes called accuracy of the confidence interval, is the distance between the estimated mean and the upper or lower range of the confidence interval.

An example of a plot of the 100(1-α)% HW confidence interval is given in Fig. 2 with 95% confidence interval in which *S*=0.08. The graph shows that beyond 10 replications there is little to be gained in terms of half width.

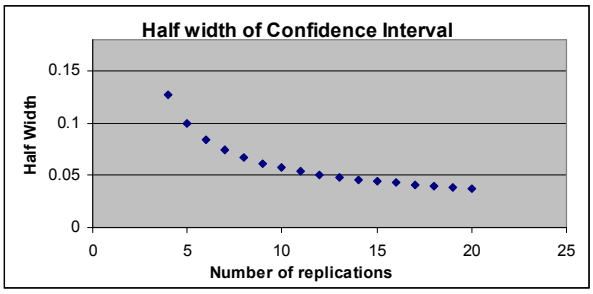


**Figure 2. 95% Half width confidence interval versus number of replications**.

The response surface and sample size techniques described in this section are the key components of a generic methodology for identifying the optimal set of GP parameters for any type of applications. The methodology can be depicted in the following diagram, shown in Figure 3.

The objective of the first step in the proposed methodology is to give an answer to the question of the necessary number of replications (independent runs) that will guarantee statistically reproducible results. The discussed approach for HW confidence interval is used and an initial number of at least 50 replicates is recommended.

The purpose of the second step is to obtain the statistically significant inputs by fractional factorial design. The next three steps are needed to find the optimal parameters. First, the new levels of the input variables (factors) which would lead to the optimum are calculated by the steepest ascent/descent method and the experimental runs on them are performed (Step 3). Second, the area around the optimum is further explored locally by a new experimental design, usually central composite design or by Box Behnken design (Step 4). The expected result from this step is a regression model around the optimum with the key factors. The final optimal parameters, however, are obtained in Step 5 by using the desirability function approach [9]. It is strongly recommended that the optimal settings are validated on other similar applications. All steps are supported by the main statistical software packages which automatically generate the corresponding experimental design plans

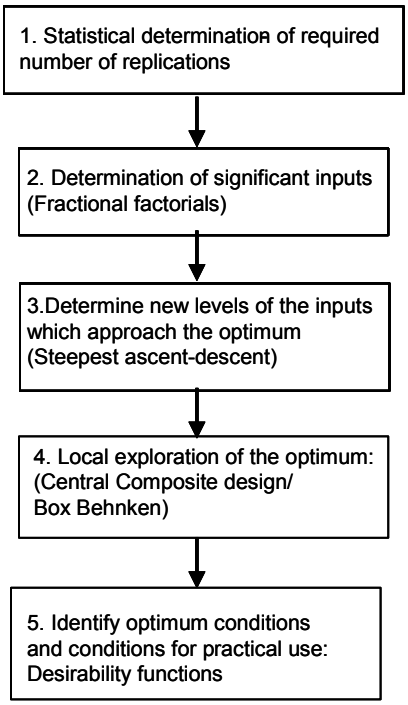


**Figure 3. Methodology for identifying the optimal set of GP parameters**

# 4. RESULTS

## 4.1 Experimental Setup

Implementing the proposed methodology requires substantial computational efforts. In order to accelerate the process, a cluster of 10 computers was used. The Pareto GP algorithm was implemented on MATLAB and the automation and facilitation of models within the cluster was done on the Legion modeling environment, an in-house developed environment. To incorporate Matlab, a python wrapper was created that would use standard functions to download, run, and then upload the simulations. In addition to this a Pipeline Pilot®[1] package protocol was created to server this through a web interface. Although Legion can run on both Linux and Windows for these simulations ten Windows compute node with 2800 MHz dual processors were used.

The full experimental study includes three types of data sets – small-sized, middle-sized, and large-sized. The small-sized data set includes up to 5 inputs and up to 50 data points. The middle-sized data set includes up to 10 inputs and up to 500 data points. The large-sized data set includes more than 10 inputs and may include thousands of data points. In this paper, the results of applying the methodology for small-sized data set will be presented. In industrial conditions, usually this type of data set is related to Design Of Experiments (DOE) for optimizing process parameters. Very often these experiments are costly and it is critical to find a solution without additional experimental efforts. The results for the other data sets and the robustness analysis will be published in a separate paper.

[1] Pipeline Pilot is a registered trademark of Sci-tegic, Accelris Inc., San Diego, CA, USA.

The small-sized data set used in the study is presented in Table 2.This small-sized data consisted of a factorial design in four variables (coded as $X_1$-$X_4$) with three center points (total of 19 experiments). In principle, the adopted coding scheme for developing models, based on DOE is as follows: the coded factors are represented with capital letters and the factors with the real measurements are represented with small letters. The factors are given in coded form with -1 as the low level, +1 as the high level, and 0 as center point. The response variable, $S_k$, was the yield or selectivity of one of the products for a chemical process [10].

**Table 2. Typical small-sized data set used to find optimum GP parameters**

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $S_k$ |
|---|---|---|---|---|
| 1 | -1 | 1 | 1 | 1.598 |
| 0 | 0 | 0 | 0 | 1.419 |
| 0 | 0 | 0 | 0 | 1.433 |
| -1 | 1 | 1 | 1 | 1.281 |
| -1 | 1 | -1 | 1 | 1.147 |
| 1 | 1 | -1 | 1 | 1.607 |
| -1 | 1 | 1 | -1 | 1.195 |
| 1 | 1 | 1 | -1 | 2.027 |
| -1 | -1 | -1 | 1 | 1.111 |
| -1 | 1 | -1 | -1 | 1.159 |
| -1 | -1 | -1 | -1 | 1.186 |
| 1 | -1 | -1 | 1 | 1.453 |
| 1 | 1 | -1 | -1 | 1.772 |
| -1 | -1 | 1 | -1 | 1.047 |
| -1 | -1 | 1 | 1 | 1.175 |
| 1 | 1 | 1 | 1 | 1.923 |
| 1 | -1 | -1 | -1 | 1.595 |
| 1 | -1 | 1 | -1 | 1.811 |
| 0 | 0 | 0 | 0 | 1.412 |

## 4.2 Optimal Pareto GP Parameters for Small-Sized Data Set

This section describes the results of applying the proposed methodology for the small-sized data set given in Table 2.

Steps 1-2. *Determination of required number of replications-Determination of significant inputs*.

In order to understand the effect of the GP inputs on the response and if the initial input settings were far removed from the optimum, a $2_v^{5-1}$ resolution V fractional factorial design was used with the inputs (GP factors) shown in Table 1 and the percentage area under the Pareto front as the response. The $2^{5-1}$ Fractional Factorial (FF) design consisted of 16 experiments plus two center points located at (30, 30, 300, 0.53, 0.75) in $x_1$, $x_2$, $x_3$, $x_4$, $x_5$ respectively. The corresponding experimental design is presented in Table 3 and was executed with 50 initial replications (independent runs). The results for the response in Table 3 are based on the average value from the 50 replications.

**Table 3. $2_v^{5-1}$ Fractional Factorial Design**

| Experimental run | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | area below Pareto front |
|---|---|---|---|---|---|---|
| 1 | 10 | 10 | 500 | 0.65 | 100 | 4.05 |
| 2 | 50 | 50 | 500 | 0.4 | 50 | 2.96 |
| 3 | 50 | 10 | 100 | 0.4 | 50 | 4.95 |
| 4 | 10 | 10 | 100 | 0.65 | 50 | 5.52 |
| 5 | 50 | 10 | 100 | 0.65 | 100 | 4.23 |
| 6 | 50 | 50 | 100 | 0.65 | 50 | 3.96 |
| 7 | 10 | 10 | 500 | 0.4 | 50 | 4.46 |
| 8 | 50 | 10 | 500 | 0.4 | 100 | 3.55 |
| 9 | 10 | 50 | 100 | 0.65 | 100 | 4.12 |
| 10 | 10 | 10 | 100 | 0.4 | 100 | 6.59 |
| 11 | 30 | 30 | 300 | 0.525 | 75 | 3.43 |
| 12 | 10 | 50 | 100 | 0.4 | 50 | 5.35 |
| 13 | 50 | 50 | 100 | 0.4 | 100 | 3.57 |
| 14 | 10 | 50 | 500 | 0.65 | 50 | 3.27 |
| 15 | 50 | 50 | 500 | 0.65 | 100 | 2.86 |
| 16 | 10 | 50 | 500 | 0.4 | 100 | 3.53 |
| 17 | 50 | 10 | 500 | 0.65 | 50 | 3.26 |
| 18 | 30 | 30 | 300 | 0.525 | 75 | 3.29 |

Calculation of the required number of replications for each experimental run was done with the 95% HW method previously described. The corresponding curves for each of the experimental runs presented in Table 3 are shown in Figure 4.

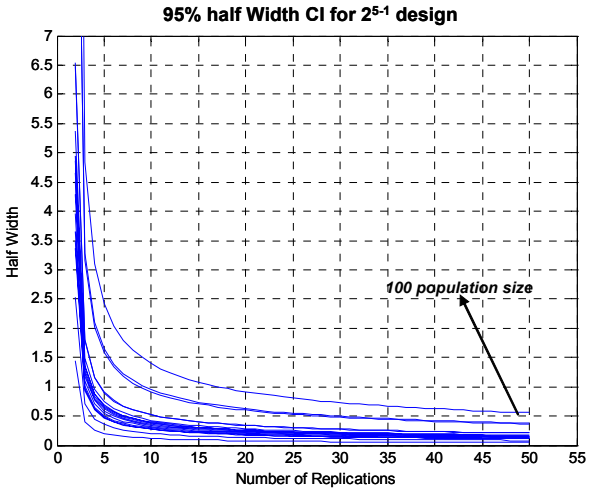


**Figure 4. A 95% Half Width Confidence interval- Determination of required number of replications for $2^{5-1}$ FF experiments**

The HW method reveals that a minimum of 50 replicates are required to be within a half width of 0.5% area under the Pareto front. Most of the experiments (except the ones for which population size was in the lower level of 100) required less than 50 replications. The average area under Pareto front is used because the variability existing between the replications for an experimental condition (the rows in Table 3) does not measure the variability between the experimental runs.

The results of the $2^{5-1}$ fractional factorial design with 50 replicates per experimental run are shown in Table 4. The Analysis was completed using the statistical package JMP®[2].

**Table 4. $2^{5-1}$ Fractional Factorial**

| **Factor** | **Estimate** | **Prob>\|t\|** |
|---|---|---|
| Intercept | 4.28 | 0.000007 |
| Number of Cascades(10,50) | -0.47 | **0.004415** |
| Number of Generations(10,50) | -0.44 | **0.006592** |
| Population Size(100,500) | -0.65 | **0.000776** |
| Prob.Func Selection(0.4,0.65) | -0.23 | 0.084668 |
| Size of Archive (50, 100) | -0.003 | 0.525436 |
| Number of Cascades*Number of Generations | 0.11 | 0.380701 |
| Number of Cascades*Population Size | 0.14 | 0.270258 |
| Number of Generations*Population Size | 0.10 | 0.411750 |
| Number of Cascades*Prob.Func Selection | 0.14 | 0.257267 |

If Prob>|t| is less than 0.05 the factor has a statistically significant effect on the response at the 95% confidence level. Based on the results, the only statistically significant factors are the number of cascades, the number of generations and the population size (highlighted in Table 4).

*Step 3.Determination of new levels of the inputs which approach the optimum*

To find conditions that led to a minimum response the path of steepest ascent-descent was calculated using the first order estimates [11]. Using the estimates of Table 4, the vector of steepest ascent is calculated as (-0.47,-.044,-0.65,-0.23, -0.003). The length of this vector is 0.93 so the unit length vector is (-0.5, -0.5, -0.7, -0.25, 0.0). Therefore of every -0.5 units in $x_1$ we need to move -0.5 in $x_2$, -0.7 in $x_3$, -0.25 in $x_4$ and 0.0 in $x_5$. The calculated path in which minimum and maximum response is expected is given in Table 5.

[2] JMP is a registered trademark of SAS Institute Inc. Cary, NC, USA.

**Table 5. Experiments in the direction of steepest ascent/descent path**

| | x1 | x2 | x3 | x4 | x5 | y |
|---|---|---|---|---|---|---|
| direction of steepest ascent | 12 | 14 | 51 | 0.47 | 75 | 6.97 |
| | 13 | 14 | 65 | 0.47 | 75 | 5.95 |
| | 15 | 16 | 92 | 0.48 | 75 | 5.21 |
| | 20 | 21 | 162 | 0.49 | 75 | 4.32 |
| Base line (center) | 30 | 30 | 300 | 0.53 | 75 | 3.3 |
| direction of steepest descent | 40 | 39 | 438 | 0.56 | 75 | 3.0 |
| | 50 | 48 | 577 | 0.59 | 75 | 2.75 |
| | 60 | 57 | 715 | 0.62 | 75 | 2.52 |
| | | | | | | |

Given that the objective is to minimize the response (percent area under Pareto front) the next experiments were planned in the direction of steepest descent centered on the base line and the results are shown in Table 5. This was decided because of the change in response around the base line (from 4.32 to 3.0) and because of practical considerations (further away from the center in the direction of steepest descent represent experiments that are unrealistic from the point of view of computation time for the required number of replications – like responses 2.75 and 2.52 in Table 5).

*Step 4. Local exploration of the optimum*

The response surface experiment uses a Box Behnken design with the significant factors shown in Table 6 (number of cascades, number of generations and population size). Table 6 gives the parameter estimates which support the regression model shown in equation (2).

**Table 6. Results from the Box Behnken Design**

| Factor | Estimate | Prob>\|t\| |
|---|---|---|
| Intercept | 3.342 | **0.000000** |
| Number of Cascades ($X_1$) | -0.354 | **0.000201** |
| Number of Generations ($X_2$) | -0.355 | **0.000198** |
| Population Size ($X_3$) | -0.503 | **0.000021** |
| Number of Cascades*Population Size | 0.227 | **0.014996** |
| Number of Cascades*Number of | 0.164 | 0.061765 |
| Number of Generations*Number of | 0.212 | **0.023804** |
| Population Size*Population Size | 0.300 | **0.004752** |

$$y = 3.342 - 0.354X_1 - 0.355X_2 - 0.503X_3 + 0.227X_1X_3 + 0.164X_1^2 + 0.212X_2^2 + 0.3X_3^2 \quad (2)$$

where $X_1$-$X_3$ are the factors shown in Table 1 in coded form between -1 and +1.

The model presented in equation (2) has an $R^2$ of 0.97 and includes statistically significant quadratic effects for number of cascades, number of generations and population size.

*Step 5. Identify optimum conditions and conditions for practical use*:

The desirability function approach [9], have then been used to find the optimal values of the parameters that minimize the response. The prediction profiler in Figure 5 below shows the setting of the parameters that minimizes the area under the Pareto front.

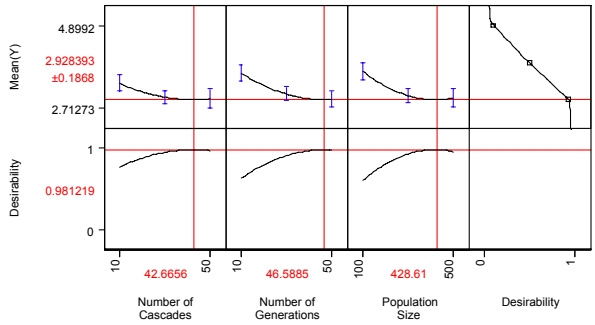


**Figure 5. Prediction Profiler showing the setting of Pareto Front GP parameters that minimizes the response.**

Of special importance are curvilinear desirability functions because they indicated a response highly sensitive to a range of input variables. As an example, in the case of number of cascades there is little improvement beyond 43. Likewise there is little improvement in minimizing the response beyond 47 for number of generations and beyond 429 for population size. These optimal values are closer to the upper range of all three statistically significant factors. For example, the range of the number of cascades is between 10 and 50 and the optimal value is 43. Table 7 depicts the optimum value for each factor for the small-sized data set.

**Table 7. Optimal Values of Pareto front GP factors**

| Factor | Optimal Value |
|---|---|
| $x_1$ - Number of cascades | 43 |
| $x_2$ - Number of generations | 47 |
| $x_3$ - Population size | 429 |
| $x_4$ - Probability of function selection | 0.53 |
| $x_5$ - Size of archive in % of pop. size | 75 |
| Number of replications | 10 |

## 5. VALIDATION

In order to verify the optimal setting of Pareto Front GP parameters, we will compare the performance with an additional small data set which is presented in Table 8 with factors coded between –1 and +1 [ 12]. The data set consisted of a Box Behnken

design on four factors ($f_1$-$f_4$) with six center points. A total of 30 experiments were performed. The response variable, y, was the particle size distribution of a chemical compound.

**Table 8. Data set for validation of optimum Pareto Front GP parameters. Box Behnken design with 6 center points**

| f1 | f2 | f3 | f4 | Y |
|---|---|---|---|---|
| -1 | -1 | 0 | 0 | 82.7 |
| -1 | 1 | 0 | 0 | 83.1 |
| 1 | -1 | 0 | 0 | 82.5 |
| 1 | 1 | 0 | 0 | 83.4 |
| 0 | 0 | -1 | -1 | 81.7 |
| 0 | 0 | -1 | 1 | 83 |
| 0 | 0 | 1 | -1 | 82.9 |
| 0 | 0 | 1 | 1 | 83.3 |
| -1 | 0 | 0 | -1 | 82.8 |
| -1 | 0 | 0 | 1 | 83.4 |
| 1 | 0 | 0 | -1 | 82.7 |
| 1 | 0 | 0 | 1 | 83.4 |
| 0 | -1 | -1 | 0 | 81.9 |
| 0 | -1 | 1 | 0 | 83 |
| 0 | 1 | -1 | 0 | 83.3 |
| 0 | 1 | 1 | 0 | 83.4 |
| -1 | 0 | -1 | 0 | 83.1 |
| -1 | 0 | 1 | 0 | 83.2 |
| 1 | 0 | -1 | 0 | 82.6 |
| 1 | 0 | 1 | 0 | 83.2 |
| 0 | -1 | 0 | -1 | 82.4 |
| 0 | -1 | 0 | 1 | 83.4 |
| 0 | 1 | 0 | -1 | 83.3 |
| 0 | 1 | 0 | 1 | 83.1 |
| 0 | 0 | 0 | 0 | 83.1 |
| 0 | 0 | 0 | 0 | 83.3 |
| 0 | 0 | 0 | 0 | 83.3 |
| 0 | 0 | 0 | 0 | 83.2 |
| 0 | 0 | 0 | 0 | 83.1 |
| 0 | 0 | 0 | 0 | 83.2 |

The optimal parameters, given in Table 7 were compared with the following non-optimal parameter set: number of cascades = 10, number of generations = 25, population size = 100, probability of function selection = 0.6, and size of the archive = 75%. The results from 10 replications (independent runs) are summarized in Table 9, where the response is the percentage of the area below the Pareto front.

**Table 9. Comparison between optimal and non-optimal response on validation data set**

| Optimal response | Non-Optimal response |
|---|---|
| 4.88 | 6.44 |
| 5.01 | 5.57 |
| 5.13 | 5.34 |
| 4.66 | 5.41 |
| 4.61 | 5.29 |
| 4.59 | 6.44 |
| 4.54 | 5.57 |
| 4.96 | 5.34 |
| 4.62 | 5.41 |
| 4.89 | 5.29 |

In order to validate if the difference in the performance based on optimal and non-optimal factors is statistically significant, a Student t test [6] is performed and the results are shown in Figure 6.

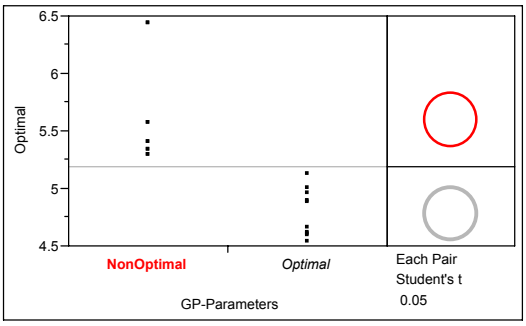


**Figure 6. Results from a Student t test on optimal and non-optimal responses for the validation data set.**

The test shows that the use of Pareto front GP optimal parameter set (given in Table 7) results in a statistically significant lower area under the Pareto front GP as compared with the non-optimal parameter set (Prob > |t|=0.00017; t ratio= 5.25). It validates the use of suggested Pareto front parameter settings on different small-sized data set and can be recommended for this class of problems.

## 6. CONCLUSION

Symbolic regression based on Pareto Front GP is the key approach for generating high-performance parsimonious empirical models acceptable for industrial applications. The paper addresses the issue of finding the optimal parameter settings of Pareto Front GP which direct the simulated evolution toward simple models with acceptable prediction error. A generic methodology based on statistical design of experiments is

proposed. It includes statistical determination of the number of replicates by half-width confidence intervals, determination of the significant inputs by fractional factorial design of experiments, approaching the optimum by steepest ascent/descent, and local exploration around the optimum by Box Behnken or by central composite design of experiments. The results from implementing the proposed methodology to a small-sized industrial data set show that the statistically significant factors for symbolic regression, based on Pareto Front GP, are the number of cascades, the number of generations, and the population size. A second order regression model with high $R^2$ of 0.97 includes the three parameters and their optimal values have been defined. The optimal parameter settings were validated with a separate small sized industrial data set. The optimal settings are recommended for symbolic regression applications using data sets with up to 5 inputs and up to 50 data points.

# 7. REFERENCES


[1] Koza, J. (1992), Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA.

[2] Banzhaf W., P. Nordin, R. Keller, and Francone, F. (1998), *Genetic Programming: An Introduction*, Morgan Kaufmann, San Francisco, CA.

[3] Feldt R. and Nordin P. (2000). Using Factorial Experiments to Evaluate the Effects of Genetic Programming parameters. In *Proceedings of EuroGP'2000*, pp. 271-282, Edinburgh, UK.

[4] Petrosky, A., Brownlee A., and McCall J. (2005), Statistical Optimization and Tuning of GA factors, In *Proceedings of the Congress of Evolutionary Computation (CEC'2005)*, pp. 758-764, Edinburgh, UK.

[5] Kordon A., F. Castillo, G. Smits, and M. Kotanchek (2006), Application Issues of Genetic Programming in Industry, In *Genetic Programming Theory and Practice III*, T. Yu, R. Riolo and B. Worzel (Eds), pp. 241-258, Springer, NY.

[6] Box, G., Hunter, W., and Hunter, J. (2005). *Statistics for Experiments: An Introduction to Design, Data Analysis, and Model Building (2nd Edition)*, New York, NY: Wiley.

[7] Smits, G. and Kotanchek, M. (2004), Pareto Front Exploitation in Symbolic Regression, In *Genetic Programming Theory and Practice II* , U.M. O'Reilly, T. Yu, R. Riolo and B. Worzel (Eds), pp 283-300, Springer, NY

[8] Montgomery, D. (1999) *Design and Analysis of Experiments*, New York, NY: Wiley

[9] Derringer, G., and Suich, R., (1980), Simultaneous Optimization of Several Response Variables, *Journal of Quality Technology*, 28(1), 61-70.

[10] Castillo, F., Marshall, K, Greens, J. and Kordon, A. (2002). Symbolic Regression in Design of Experiments: A Case Study with Linearizing Transformations, In *Proceedings of the Genetic and Evolutionary Computing Conference (GECCO'2002)*, W. Langdon, *et al* (Eds), pp. 1043-1048. New York, NY: Morgan Kaufmann.

[11] Myers, R., Montgomery, D., (1995), *Response Surface Methodology- Process and Product Optimization using Design of Experiments*, John Wiley & Sons, New York

[12] Castillo, F., Sweeney, J., and Zirk, W.(2004), Using Evolutionary Algorithms to Suggest Variable Transformations in Linear Model Lack-of-Fit Situations, in *Proceedings of the Congress of Evolutionary Computation (CEC'2004)*, pp. 556-560, Portland, OR.