

# Practical experience of integrating Genetic Programming and statistical modeling

Flor Castillo

The Dow Chemical Company  
Freeport, TX, USA

# References

- Castillo, F., Sweeney, J., Zirk, W. (2004), Using Evolutionary Algorithms to Suggest Variable Transformations in Linear Model Lack-of-Fit Situations, Congress on Evolutionary Computation, 556-568.
- Castillo, F., Kordon, A., Sweeney, J., Zirk, W. (2004), Using Genetic Programming in Industrial Statistical Model Building, in the book Genetic Programming Theory and Practice, Boston, MA. 31-47.
- Castillo, F., Kordon A., Smits, G. (2007), Robust Pareto Front Genetic Programming Parameter Selection Based on Design of Experiments and Industrial Data, published in the book Genetic Programming Theory and Practice, Springer, 150-165

# Outline

- Synergy between Genetic Programming (GP) & statistical model building
- GP applied to Industrial Statistical Modeling
  - GP applied to designed data
  - GP applied to undesigned data
- Statistical techniques applied to GP

# Unique Features of GP Attractive to Statistical Modeling

$$GPFunction1 = e^{-x_7} - \text{Log}[-\text{Log}[x_6^2] - x_4^2 + x_5 + x_7]^2 - \sqrt{x_2 + x_2 + x_4}$$

$$S_k = \frac{3.13868 \times 10^{-17} e^{\sqrt{2x_1}} \ln[(x_3)^2] x_2 + 1.00545}{x_4}$$

$$y = a + b \left( \frac{\sqrt{\frac{e^{-x_3}}{\log(x_1 x_5^2)}}}{e^{-x_3} + \log(x_2)} + \sqrt{x_1} + x_5 \right) \quad (2)$$

**GP Features**

Diverse model generation

No variable independence assumption

No least-squares estimation assumptions

Errors are normally distributed

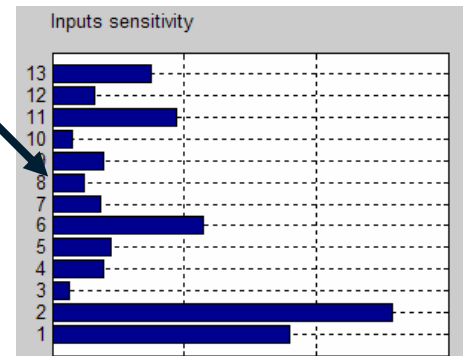
Errors with zero mean

Errors with constant variance

Nonlinear transforms that can linearize the problem

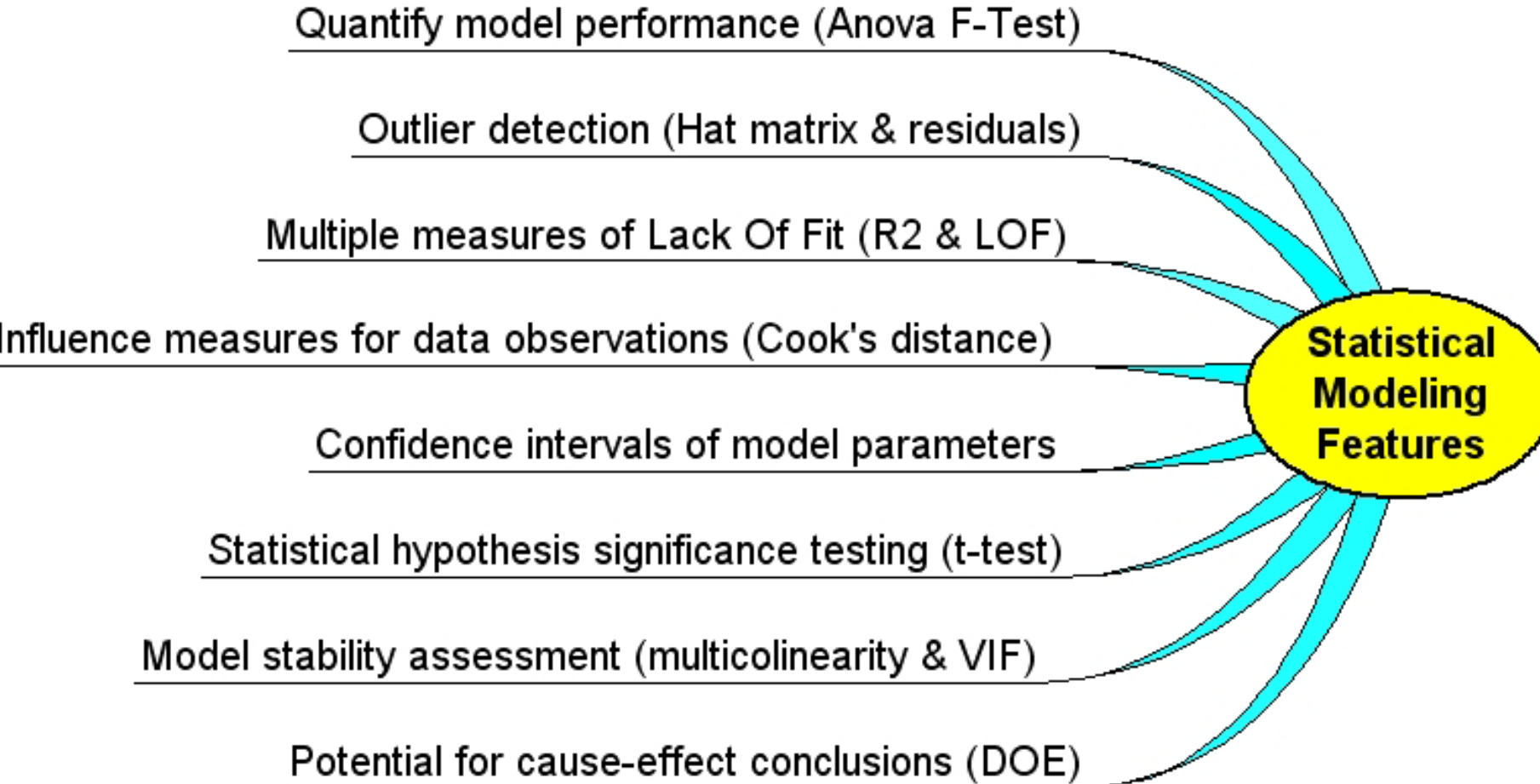
Sensitivity analysis and variable selection

Model generation from small data sets

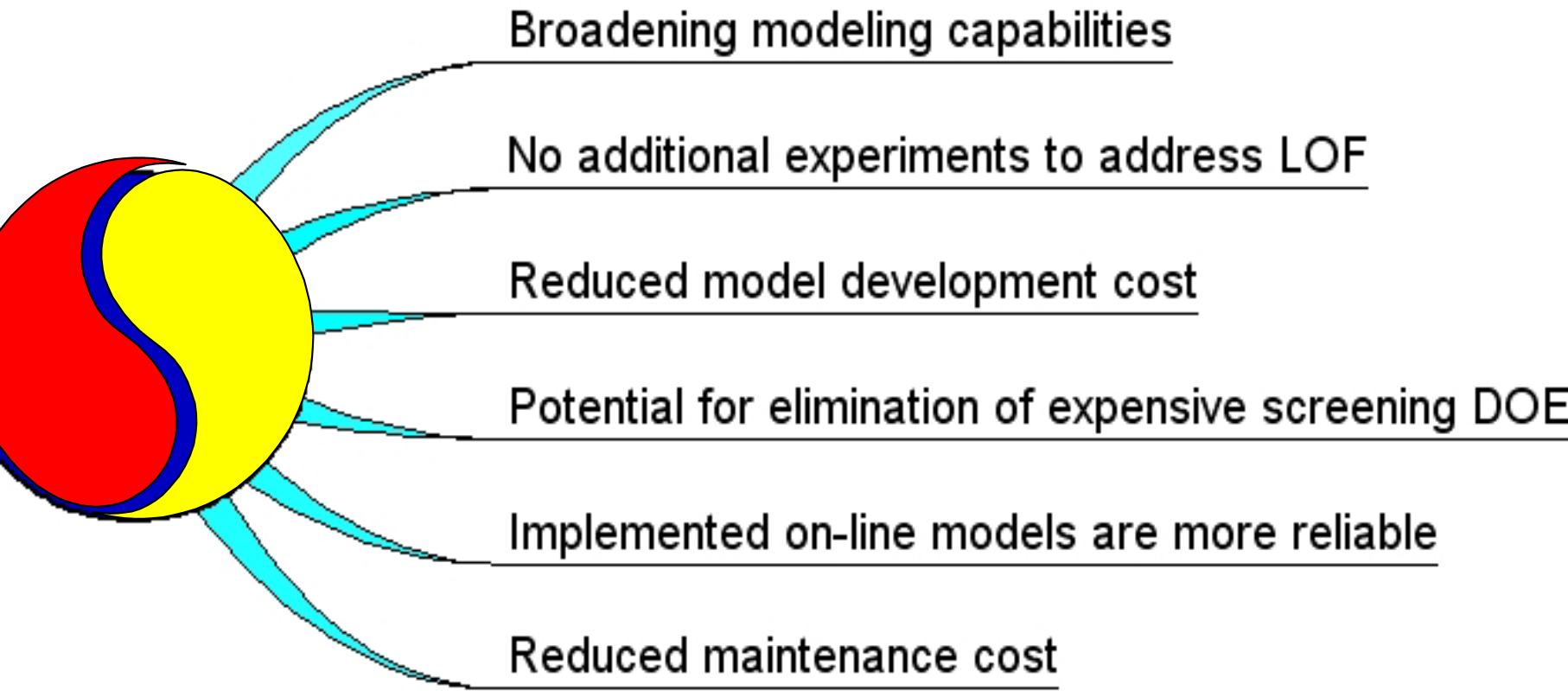


# Unique Features of Statistical Modeling GP

## Attractive to GP



# Synergetic benefits



# 1. Using GP in Industrial Statistical Model Building

## Two case studies

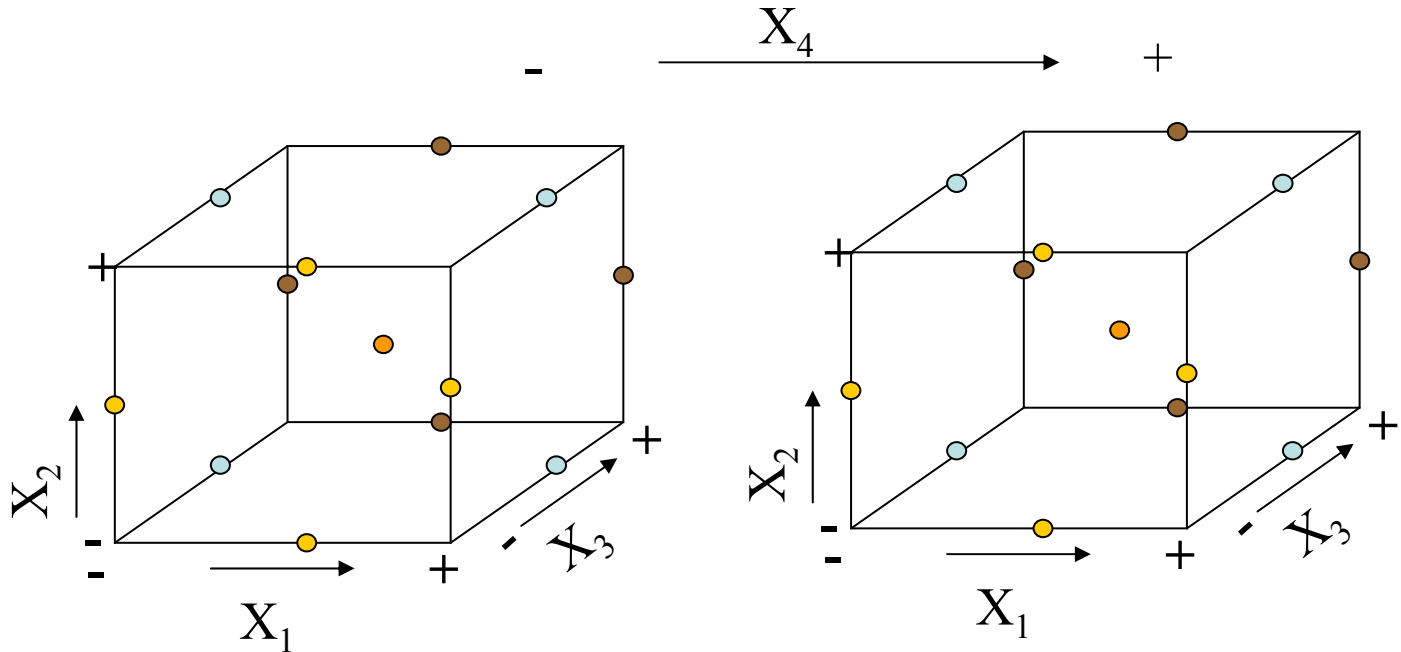
- “Designed” data collected using statistical design of experiments (DOE) scheme
  - Broad process changes made to understand variable cause & effect
- “Undesigned” (historical) data collected over time as naturally occurs
  - Observational data typically collected for control purposes

# Designed Data Case Study

- In a chemical process, experimental data obtained for process improvement purposes
- Particle size distribution of chemical compound is response (output) of interest
- Four input variable levels varied
- Box-Behnken response surface experimental design plan completed



# Box-Behnken Experimental Design



## Response (Output):

Particle size distribution of  
a chemical compound

## Inputs:

•  $X_1, X_2, X_3, X_4$

$$S_k = \beta_o + \sum_{i=1}^k \beta_i X_i + \sum_{i < j} \sum \beta_{ij} X_i X_j + \sum \beta_{ii} X_i^2$$

# Box-Behnken Data Analysis

**Full Model**  
 $R^2 = 0.88$

**Reduced model (without  $X_1$  terms)**  
 $R^2 = 0.85$

**Analysis of Variance**

| Source   | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|----------|----|----------------|-------------|---------|----------|
| Model    | 14 | 4.711          | 0.336       | 7.78    |          |
| Error    | 15 | 0.649          | 0.043       |         |          |
| C. Total | 29 | 5.360          |             |         | 0.0002   |

**Lack Of Fit**

| Source      | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|-------------|----|----------------|-------------|---------|----------|
| Lack Of Fit | 10 | 0.609          | 0.061       | 7.61    |          |
| Pure Error  | 5  | 0.040          | 0.008       |         |          |
| Total Error | 15 | 0.649          |             |         | 0.0185   |

Max RSq  
0.993

**Parameter Estimates**

| Term            | Estimate | t Ratio | Prob> t |
|-----------------|----------|---------|---------|
| Intercept       | 83.2     | 979.64  | <.0001  |
| X1(700,2100)&RS | -0.0417  | -0.69   | 0.4984  |
| X2(20,40)&RS    | 0.30833  | 5.13    | 0.0001  |
| X3(3,15)&RS     | 0.28333  | 4.72    | 0.0003  |
| X4(8,16)&RS     | 0.31667  | 5.27    | <.0001  |
| X1*X1           | -0.0375  | -0.47   | 0.6437  |
| X2*X1           | 0.125    | 1.20    | 0.2481  |
| X2*X2           | -0.1125  | -1.42   | 0.1772  |
| X3*X1           | 0.125    | 1.20    | 0.2481  |
| X3*X2           | -0.25    | -2.40   | 0.0296  |
| X3*X3           | -0.225   | -2.83   | 0.0126  |
| X4*X1           | 0.025    | 0.24    | 0.8133  |
| X4*X2           | -0.3     | -2.88   | 0.0114  |
| X4*X3           | -0.225   | -2.16   | 0.0471  |
| X4*X4           | -0.125   | -1.57   | 0.1365  |

**Significant Lack-of-fit in full model**

**All Terms involving  $X_1$  are not significant**

**Analysis of Variance**

| Source   | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|----------|----|----------------|-------------|---------|----------|
| Model    | 9  | 4.553          | 0.506       | 12.53   |          |
| Error    | 20 | 0.807          | 0.040       |         |          |
| C. Total | 29 | 5.360          |             |         | <.0001   |

**Lack Of Fit**

| Source      | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|-------------|----|----------------|-------------|---------|----------|
| Lack Of Fit | 9  | 0.572          | 0.064       | 2.98    |          |
| Pure Error  | 11 | 0.235          | 0.021       |         |          |
| Total Error | 20 | 0.807          |             |         | 0.0460   |

Max RSq  
0.956

**Parameter Estimates**

| Term         | Estimate | t Ratio | Prob> t |
|--------------|----------|---------|---------|
| Intercept    | 83.2     | 979.64  | <.0001  |
| X2(20,40)&RS | 0.30833  | 5.13    | 0.0001  |
| X3(3,15)&RS  | 0.28333  | 4.72    | 0.0003  |
| X4(8,16)&RS  | 0.31667  | 5.27    | <.0001  |
| X2*X2        | -0.1125  | -1.42   | 0.1772  |
| X3*X2        | -0.25    | -2.40   | 0.0296  |
| X3*X3        | -0.21964 | -2.89   | 0.0090  |
| X4*X2        | -0.3     | -2.99   | 0.0073  |
| X4*X3        | -0.225   | -2.24   | 0.0366  |
| X4*X4        | -0.11964 | -1.58   | 0.1308  |

**Still, significant Lack-of-fit in reduced model**

# Lack of Fit Situations

## 1. Two-level factorial DOE

Experimental data available to fit first order model only

$$S_k = \beta_o + \sum_{i=1}^k \beta_i X_i + \sum_{i < j} \beta_{ij} X_i X_j$$

**What if LOF is present?**

 **Classical approach: add experiments to enable second order model fit**


$$S_k = \beta_o + \sum_{i=1}^k \beta_i X_i + \sum_{i < j} \beta_{ij} X_i X_j + \sum \beta_{ii} X_i^2$$

## 2. Response surface DOE

Experimental data available to fit second order model

$$S_k = \beta_o + \sum_{i=1}^k \beta_i X_i + \sum_{i < j} \beta_{ij} X_i X_j + \sum \beta_{ii} X_i^2$$

**What if LOF is present?**

 **No obvious single approach recommended ...**



**What can we do?**

# Possible LOF Solutions

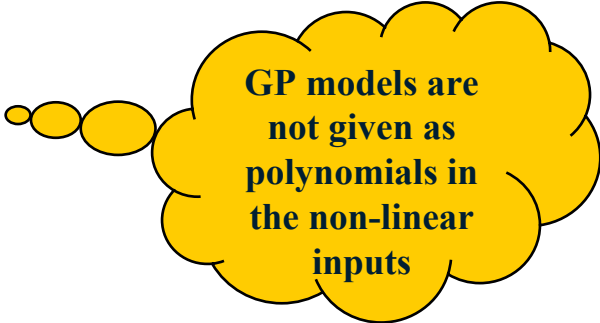
- Ignore it
  - Possible limitations on conclusions
- Collect more data
  - Induce correlation
  - Cost of additional sampling, etc.
- Try a different more complex model
  - Current data may not support new model
- Try a different transformed model
  - Transformation to try not obvious (**can Genetic Programming (GP) help?**)

# GP Algorithm Details

- GP algorithm uses many mathematical operators to assemble expressions
  - Addition, subtraction, multiplication, division, square, square root, change sign, natural logarithm, exponential, power, & numerical constants
- Each expression evaluated using fitness function
  - Typically, SSE or correlation coefficient
- High fitness functions passed to next generation & are combined (crossover) to try new expressions
  - Sometimes a new expression (mutation) introduced
  - Best equations have higher chance of inclusion in next generation
- Process repeats many times until user-defined criteria are completed
  - Number of generations & number of runs

# Suggested LOF Process

1. Generate several GP models
2. Generate linear model in the non-linear input transforms
3. Fit response surface model in transformed variables
4. Perform statistical analysis (check LOF, error structure, residuals, correlations (VIF))
5. Model discrimination



**GP models are not given as polynomials in the non-linear inputs**

# LOF Process – Step 1

1. Generate several GP models
2. Generate linear model in the non-linear input transforms
3. Fit response surface model in transformed variables
4. Perform statistical analysis (check LOF, error structure, residuals, correlations (VIF))
5. Model discrimination

$$\text{solution 1} = \frac{1}{x_2^2 x_3 x_4^2}$$

$$\text{solution 2} = \sqrt{\text{Abs}\left[\text{Log}\left[-9.327 - x_2 - x_3\right] - \frac{3.6773}{1 - x_3 + \frac{x_2}{x_4} - x_4}\right]}$$

$$\text{solution 3} = \text{Abs}\left[\text{Log}\left[e^{-x_4} + \frac{9.327}{x_2}\right] - \frac{3.6773}{-x_3 + \frac{2x_2}{x_4} - x_4}\right]^{1/4}$$

$$\text{solution 4} = -\frac{1}{x_3 x_4}$$

# LOF Process – Step 2

1. Generate several GP models
2. Generate linear model in the non-linear input transforms
3. Fit response surface model in transformed variables
4. Perform statistical analysis (check LOF, error structure, residuals, correlations (VIF))
5. Model discrimination

$$y = \frac{|x_2|^{0.54528}}{\sqrt{\ln(x_3 x_2 + x_3)}} * x_2 x_4$$

Notice  
solution does  
not involve  
 $X_1$

**Variable transformations suggested by GP model**

| Original Variable | Transformed Variable             |
|-------------------|----------------------------------|
| $X_2$             | $Z_2 = X_2^{0.5}$                |
| $X_3$             | $Z_3 = [\text{Log}(X_3)]^{-0.5}$ |
| $X_4$             | $Z_4 = X_4^{-1}$                 |

# LOF Process – Step 3

1. Generate several GP models
2. Generate linear model in the non-linear input transforms
3. Fit response surface model in transformed variables
4. Perform statistical analysis (check LOF, error structure, residuals, correlations (VIF))
5. Model discrimination

$$S_k = \beta_o + \sum_{i=2}^4 \beta_i Z_i + \sum_{i < j} \beta_{ij} Z_i Z_j + \sum_{i=2}^4 \beta_{ii} Z_i^2$$

| Parameter Estimates    |          |         |         |       |
|------------------------|----------|---------|---------|-------|
| Term                   | Estimate | t Ratio | Prob> t | VIF   |
| Intercept              | 82.8704  | 748.58  | <.0001  | .     |
| Z2(4.47214,6.32456)&RS | 0.4771   | 7.31    | <.0001  | 1.573 |
| Z3(0.60768,0.95406)&RS | -0.3578  | -6.49   | <.0001  | 1.371 |
| Z4(0.0625,0.125)&RS    | -0.4379  | -7.14   | <.0001  | 1.477 |
| Z2*Z2                  | -0.0887  | -1.29   | 0.2128  | 1.034 |
| Z2*Z3                  | 0.28248  | 3.50    | 0.0022  | 1.415 |
| Z3*Z3                  | -0.0724  | -0.60   | 0.5556  | 1.254 |
| Z2*Z4                  | 0.23959  | 2.75    | 0.0123  | 1.151 |
| Z3*Z4                  | -0.2631  | -3.37   | 0.0030  | 1.525 |
| Z4*Z4                  | -0.0166  | -0.21   | 0.8362  | 1.095 |

# LOF Process – Step 4

1. Generate several GP models
2. Generate linear model in the non-linear input transforms
3. Fit response surface model in transformed variables
4. Perform statistical analysis (check LOF, error structure, residuals, correlations (VIF))
5. Model discrimination

## Analysis of Variance

| Source   | DF | Sum of Squares | Mean Square |
|----------|----|----------------|-------------|
| Model    | 9  | 4.7080         |             |
| Error    | 20 | 0.6520         |             |
| C. Total | 29 | 5.3600         |             |

Notice no Lack-of-fit

## Lack Of Fit

| Source      | DF | Sum of Squares | Mean Square | Prob>F |
|-------------|----|----------------|-------------|--------|
| Lack Of Fit | 9  | 0.4170         | 0.0463      | 0.59   |
| Pure Error  | 11 | 0.2350         | 0.0214      | 0.1131 |
| Total Error | 20 | 0.6520         |             |        |

Max RSq  
0.956

## Parameter Estimates

| Term                   | Estimate | t Ratio | Prob> t | VIF   |
|------------------------|----------|---------|---------|-------|
| Intercept              | 82.8704  | 748.58  | <.0001  | .     |
| Z2(4.47214,6.32456)&RS | 0.4771   | 7.31    | <.0001  | 1.573 |
| Z3(0.60768,0.95406)&RS | -0.2552  | -6.49   | <.0001  | 1.371 |
| Z4(0.00000,0.00000)&RS | 0.0000   | 0.00    | 0.9999  | 1.477 |
| Z2^2                   | 0.0000   | 0.2128  | 0.8362  | 1.034 |
| Z3^2                   | 0.0000   | 0.0022  | 0.9978  | 1.415 |
| Z4^2                   | 0.0000   | 0.0000  | 0.9999  | 1.254 |
| Z2*Z3                  | 0.0000   | 0.60    | 0.5556  | 1.254 |
| Z2*Z4                  | 0.0000   | 2.75    | 0.0123  | 1.151 |
| Z3*Z4                  | -0.2631  | -3.37   | 0.0030  | 1.525 |
| Z4*Z4                  | -0.0166  | -0.21   | 0.8362  | 1.095 |

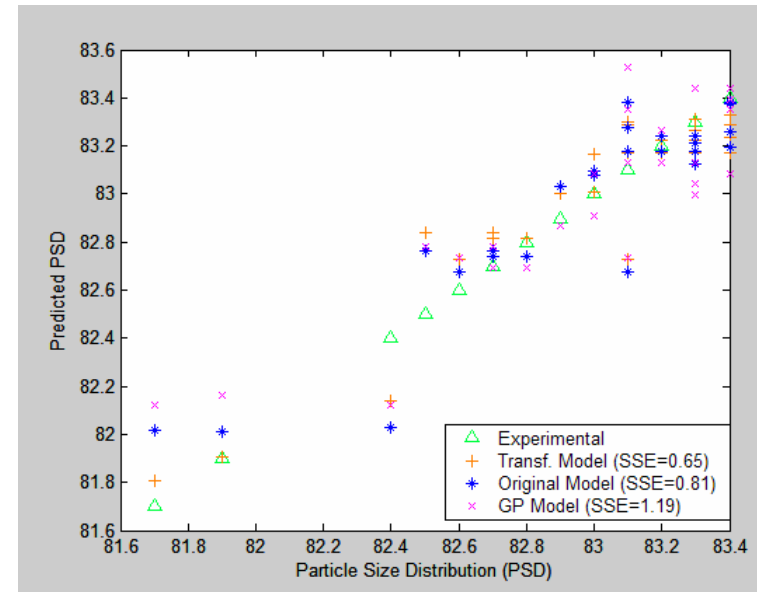
Collinearity is no problem (VIF's < 8.33)

## Summary of Fit

|                           |       |
|---------------------------|-------|
| RSquare                   | 0.878 |
| RSquare Adj               | 0.824 |
| Root Mean Square Error    | 0.181 |
| Mean of Response          | 83    |
| Observations (or Sum Wts) | 30    |

# LOF Process – Step 5

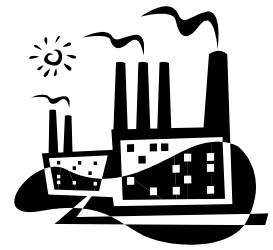
1. Generate several GP models
2. Generate linear model in the non-linear input transforms
3. Fit response surface model in transformed variables
4. Perform statistical analysis (check LOF, error structure, residuals, correlations (VIF))
5. Model discrimination



**PSD Model Comparison**

|                   | $R^2$ | SSE  | Model parameters | LOF   |
|-------------------|-------|------|------------------|-------|
| Original (RSM)    | 0.85  | 0.81 | 9                | 0.046 |
| Transf. Model     | 0.88  | 0.65 | 9                | 0.113 |
| GP model (Genpro) | 0.77  | 1.19 | na               | na    |
|                   |       |      |                  |       |

# Undesigned Data Case Study



- In another chemical process, data obtained from 3-month process history was used in empirical modeling effort
- A (detrimental) by-product concentration was response (output) of interest
- All other variables considered potential inputs
- Can a reasonable empirical model be developed to predict how this by-product output can be minimized?

# Regression Model Analysis

- Undesigned data will often be too unbalanced for standard modeling techniques
- Variance Inflation Factors (VIF) often used to identify potential data imbalance (multicollinearity) issues
- Large VIF's suggest problems
- In this case, VIF's suggest that regression model is very unstable!

| Term      | $\beta$ Estimate | t Ratio | Prob> t | VIF       |
|-----------|------------------|---------|---------|-----------|
| Intercept | 230.70902        | 0.33    | 0.7432  |           |
| x1        | 0.9406677        | 19.31   | <0.0001 | 3.84056   |
| x2        | -2.428614        | -22.97  | <0.0001 | 7.05279   |
| x3        | 0.4005954        | 2.97    | 0.0041  | 9.42801   |
| x4        | -10.17105        | -0.36   | 0.7217  | 861.2503  |
| x5        | 2.956458         | 0.20    | 0.8385  | 343.7906  |
| x6        | 10.223555        | 0.36    | 0.7164  | 918.9986  |
| x7        | -31.91927        | -0.57   | 0.5686  | 3431.5002 |
| x8        | 14.871442        | 0.35    | 0.7257  | 1976.0583 |
| x9        | -135.1481        | -0.69   | 0.4919  | 1000231.8 |
| x10       | 117.8077         | 0.68    | 0.4967  | 964097.17 |
| x11       | 16.152238        | 0.40    | 0.6930  | 70850.669 |
| x12       | 14.186557        | 0.89    | 0.3750  | 77.489476 |
| x13       | -19.53814        | -0.67   | 0.5023  | 19404.123 |

**Can GP suggest an alternative model to try?**

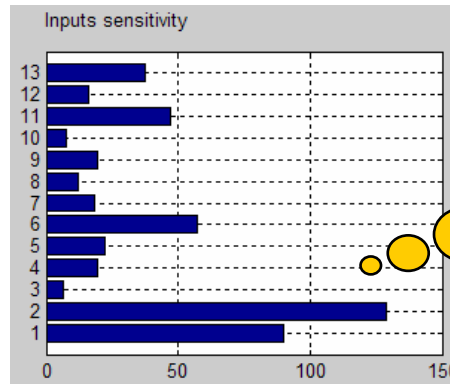
# GP Algorithm Results

## What alternative models does GP suggest?

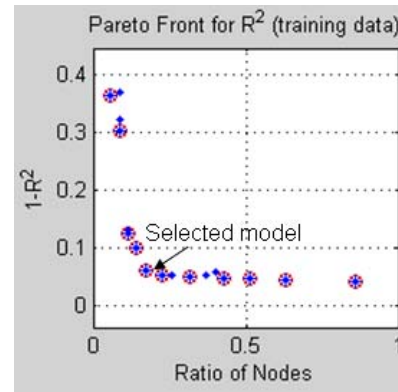
Algorithm parameters used:

- 50 runs
- 20 generations
- Population size of 100

Pareto front optimization used to select model with “best” balance between performance & complexity



$X_1, X_2, X_6, X_{11},$  &  $X_{13}$  were included most often



$$y = 10275 - 16.078 \cdot \frac{x_6(x_2 + x_{11})}{x_1 + x_{13}}$$

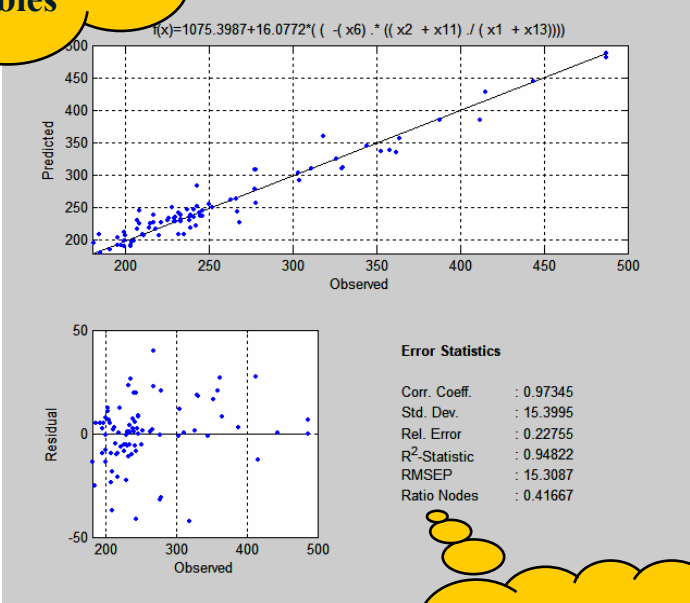
Selected model

# GP Generated Transformed Model

$$Y = 10275 - 16078 \frac{x_6(x_2 + x_{11})}{x_1 + x_{13}}$$

Selected GP model suggests new transformed variables

| Original Variable | Transformed Variable       |
|-------------------|----------------------------|
| $x_2, x_{11}$     | $Z_1 = (x_2 + x_{11})$     |
| $x_1, x_{13}$     | $Z_2 = 1 / (x_1 + x_{13})$ |
| $x_6$             | $Z_3 = x_6$                |



| Term                   | $\beta$ Estimate | t Ratio | Prob> t | VIF   |
|------------------------|------------------|---------|---------|-------|
| Intercept              | 2955.597         | 16.616  | <0.0001 |       |
| $Z_3 = x_6$            | -7.265           | -5.812  | <0.0001 | 1.496 |
| $Z_1 = x_2 + x_{11}$   | -2.148           | -32.646 | <0.0001 | 2.504 |
| $Z_2 = 1/x_1 + x_{13}$ | -908023.43       | -21.148 | <0.0001 | 2.392 |

No collinearity problems

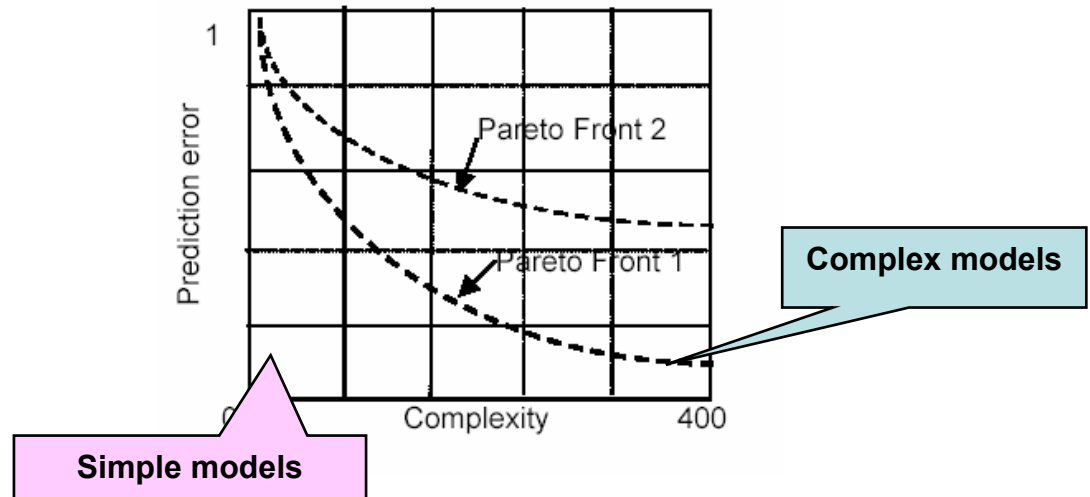
Transformed model looks promising

# Advantages of GP For Designed and Undesigned Data

- All well-developed statistical techniques applied for model discrimination
- GP suggests transformations that normally would not be considered from conventional Taylor series expansion models
- GP suggested transform eliminates model lack of fit without requiring additional experimental data
- Transformed model appears to be more stable than original model

## 2. Using Statistical Techniques to Improve GP

- Select GP parameters to push the simulated evolution to the lower left Pareto Front corner



# Key Pareto Front GP Parameters

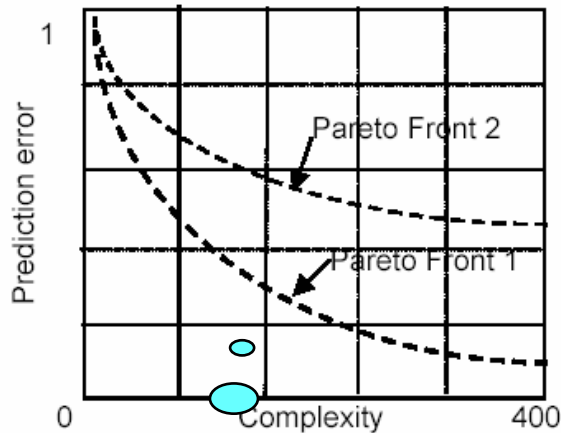
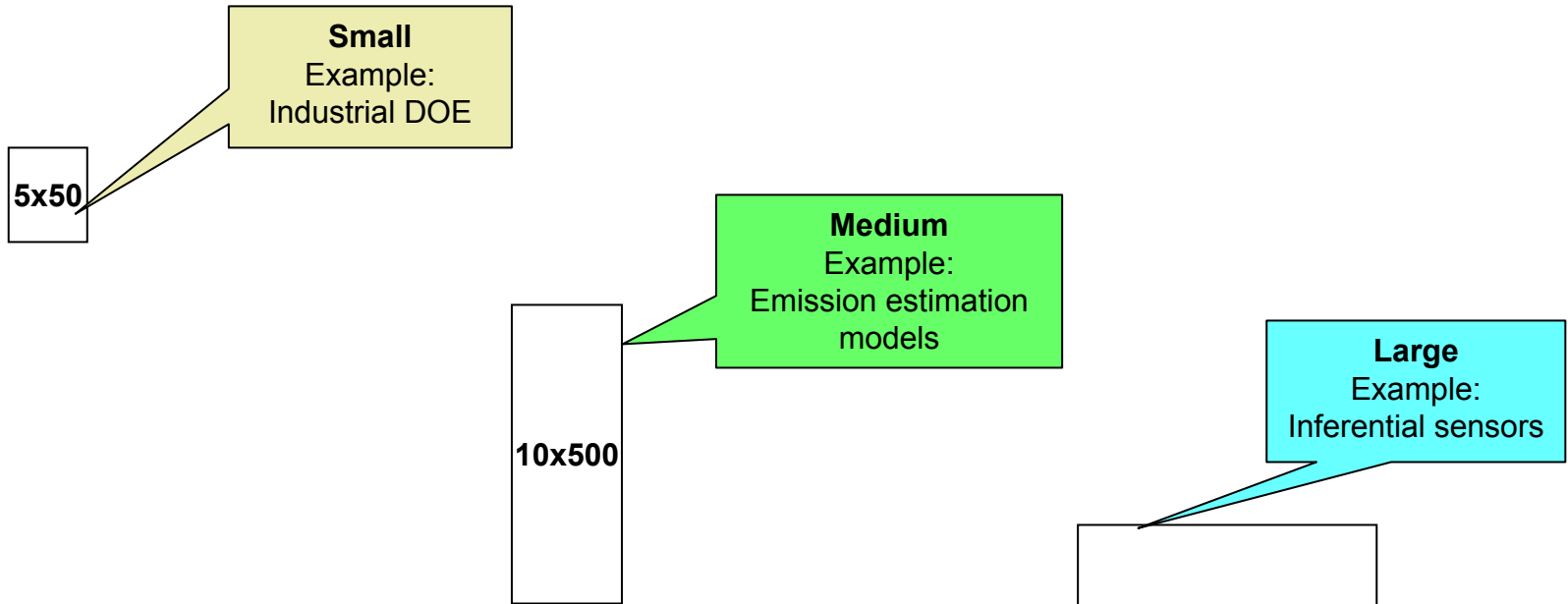


Table 1. Factors for the Pareto Front GP DOE

| 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             |

- Independent run (replicate): begins with random population in archives
- Cascade: Pareto Front archive maintained
- Experimental run: uses specific parameter set, defined by the DOE

# Typical Industrial Data Sets for Empirical Modeling



**Key issue:**  
How robust is parameter selection  
toward different data sets?

# RSM Pareto-Front GP Parameters- Step 1

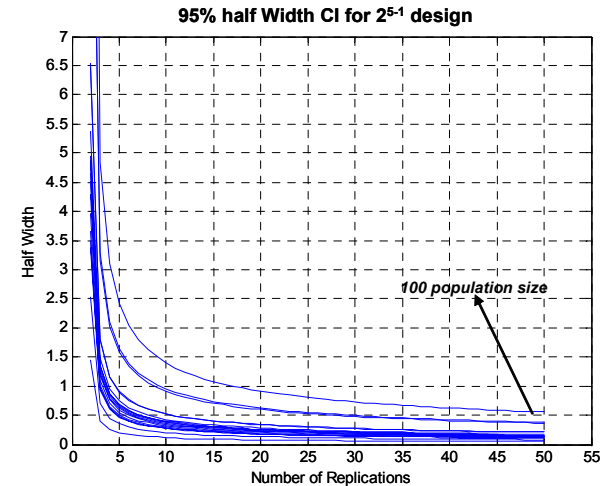
1. Statistical determination of required number of replications

2. Determination of significant inputs (Fractional factorials)

3. Determine new levels of the inputs which approach the optimum (Steepest ascent-descent)

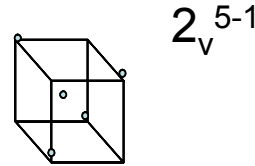
4. Local exploration of the optimum:(Central Composite design/Box Behnken)

5. Identify optimum conditions and conditions for practical use:Desirability functions

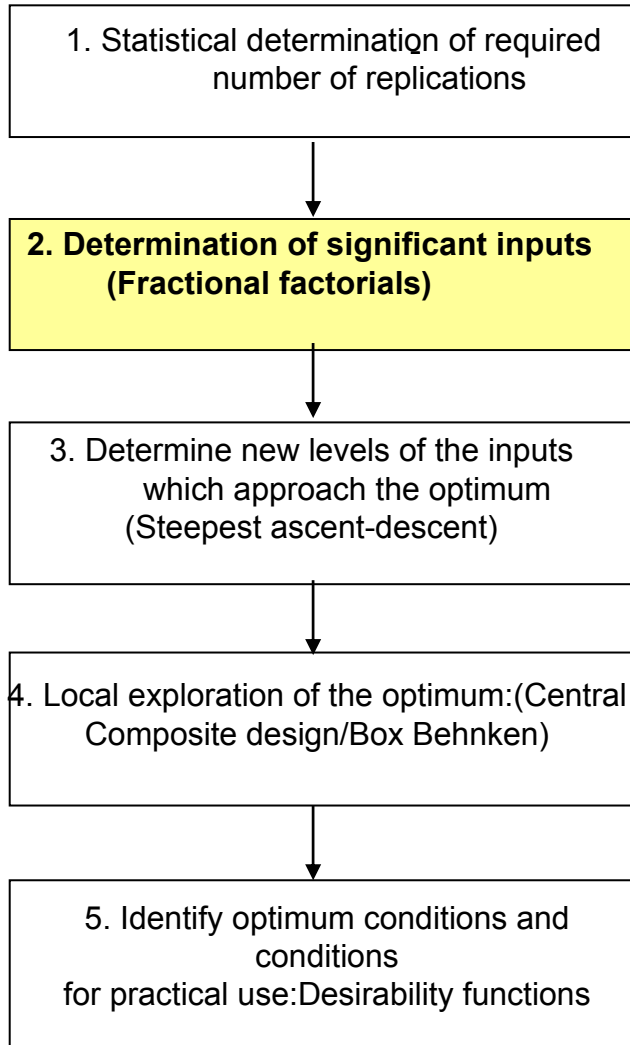


$$HW = t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$

# RSM Pareto-Front GP Parameters- Step 2

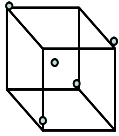


17 runs instead of 32



| Experimental run | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | Response |
|------------------|-------|-------|-------|-------|-------|----------|
| 1                | 10    | 10    | 500   | 0.65  | 100   | 6.02     |
| 2                | 50    | 50    | 500   | 0.4   | 50    | 5.38     |
| 3                | 50    | 10    | 100   | 0.4   | 50    | 6.17     |
| 4                | 10    | 10    | 100   | 0.65  | 50    | 6.69     |
| 5                | 50    | 10    | 100   | 0.65  | 100   | 6.00     |
| 6                | 50    | 50    | 100   | 0.65  | 50    | 5.94     |
| 7                | 10    | 10    | 500   | 0.4   | 50    | 6.02     |
| 8                | 50    | 10    | 500   | 0.4   | 100   | 5.69     |
| 9                | 10    | 50    | 100   | 0.65  | 100   | 5.99     |
| 10               | 10    | 10    | 100   | 0.4   | 100   | 7.88     |
| 11               | 30    | 30    | 300   | 0.525 | 75    | 5.58     |
| 12               | 10    | 50    | 100   | 0.4   | 50    | 6.28     |
| 13               | 50    | 50    | 100   | 0.4   | 100   | 5.70     |
| 14               | 10    | 50    | 500   | 0.65  | 50    | 5.58     |
| 15               | 50    | 50    | 500   | 0.65  | 100   | 5.23     |
| 16               | 10    | 50    | 500   | 0.4   | 100   | 5.68     |
| 17               | 50    | 10    | 500   | 0.65  | 50    | 5.53     |

# RSM Pareto-Front GP Parameters- Step 2



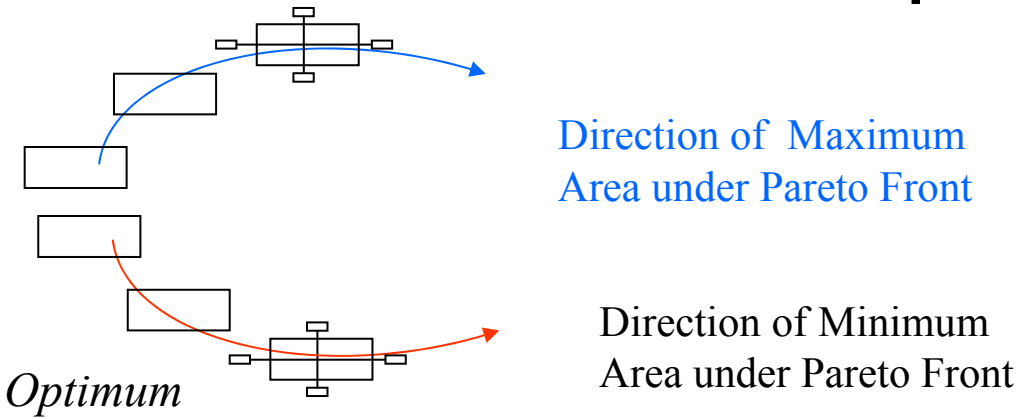
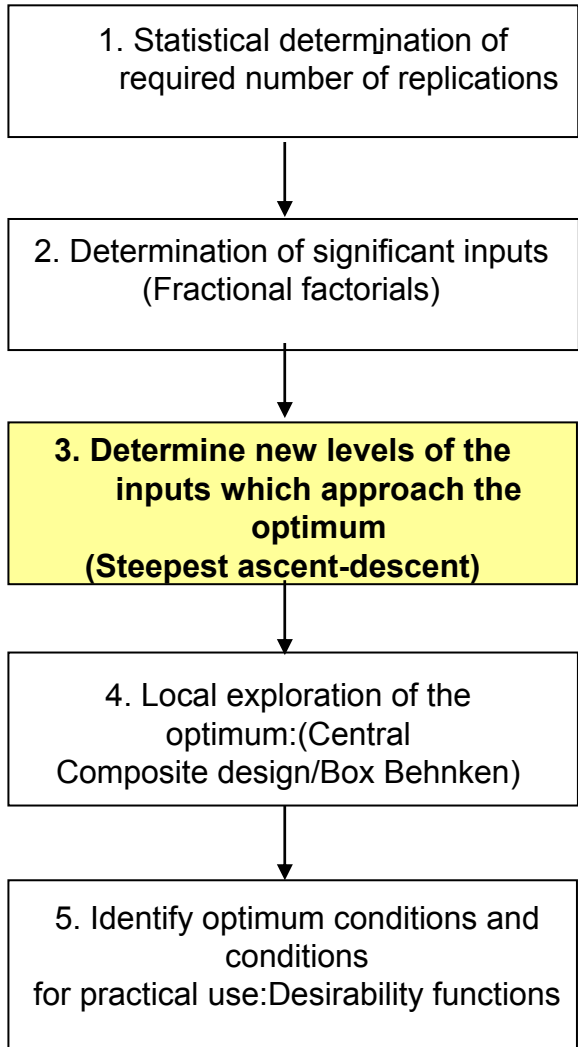
Statistical Results from  $2_{V^{5-1}}$  Fractional Factorial Design for a Medium-Sized Data Set

| Factor                                   | Estimate | Prob> t         |
|--|----------|-----------------|
| Intercept                                | 6.27     | 4.53E-05        |
| <b>Number of Cascades</b>                | -0.28    | <b>0.011958</b> |
| <b>Number of Generations</b>             | -0.26    | <b>0.016473</b> |
| <b>Population Size</b>                   | -0.34    | <b>0.004191</b> |
| Prob.Func Selection                      | -0.11    | 0.229607        |
| Size of Archive                          | 0.29     | 0.686794        |
| Number of Cascades*Number of Generations | 0.12     | 0.200071        |
| Number of Cascades*Population Size       | 0.10     | 0.288629        |
| Number of Generations*Population Size    | 0.09     | 0.333296        |

(-0.28,-0.26,-0.34,-0.11, 0.29).

statistically significant inputs  
Prob>|t| < 0.05

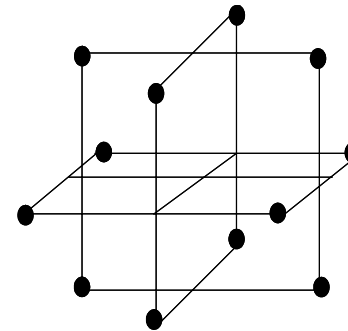
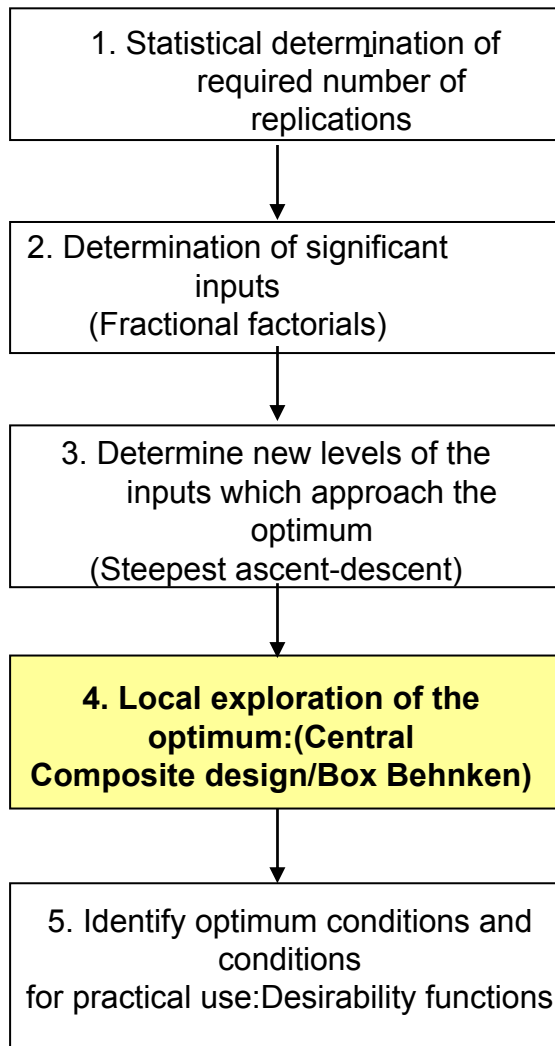
# RSM Pareto-Front GP Parameters- Step 3



Vector:  $(-0.28, -0.26, -0.34, -0.11, 0.29)$ .  
 unit length vector  $(-0.47, -0.4, -0.57, -0.19, 0.48)$ .  
 for every  $-0.47$  units in  $X1$  move  $-0.4$  in  $X2$ ,  
 $-0.57$  in  $X3$ ,  $-0.19$  in  $X4$  and  $0.48$  in  $X5$

| Direction        | x1 | x2 | x3  | x4   | x5 | y    |
|------------------|----|----|-----|------|----|------|
| Steepest ascent  | 13 | 14 | 95  | 0.48 | 97 | 6.51 |
|                  | 14 | 15 | 106 | 0.49 | 95 | 6.23 |
|                  | 16 | 17 | 129 | 0.49 | 93 | 6.11 |
|                  | 21 | 21 | 186 | 0.5  | 87 | 5.86 |
| Base line        | 30 | 30 | 300 | 0.53 | 75 | 5.58 |
| Steepest descent | 39 | 39 | 414 | 0.55 | 63 | 5.38 |
|                  | 49 | 47 | 528 | 0.57 | 51 | 5.30 |

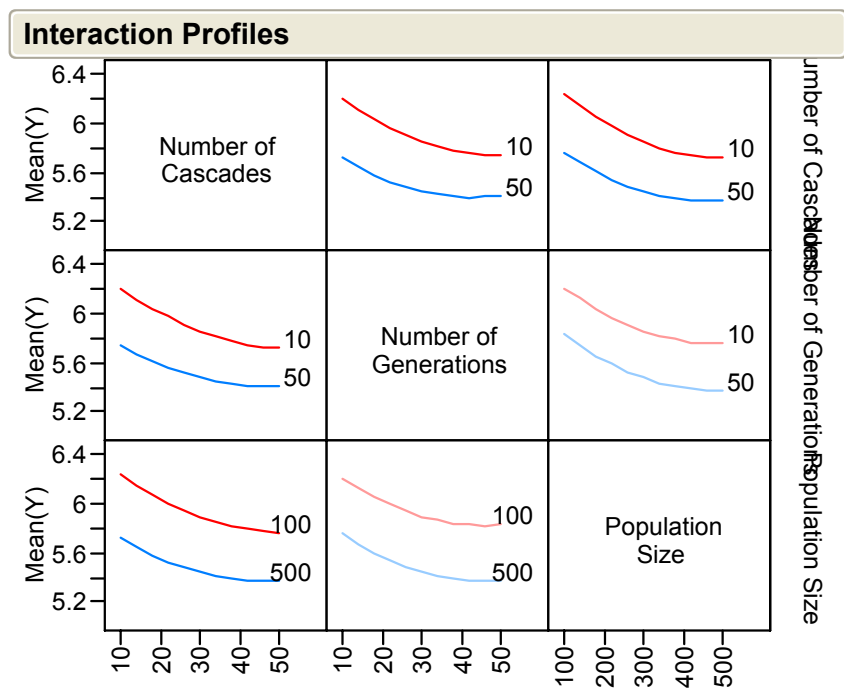
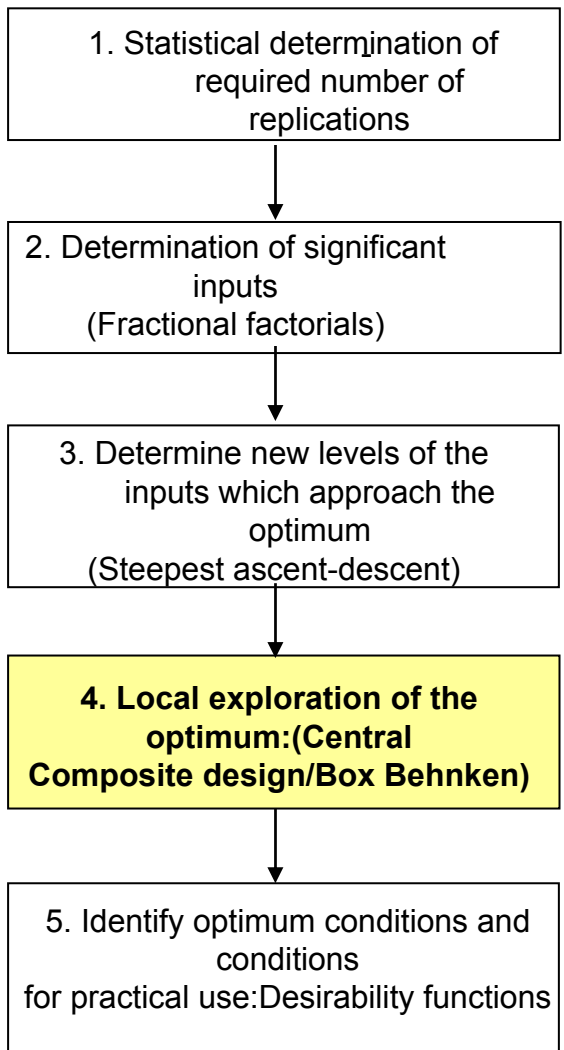
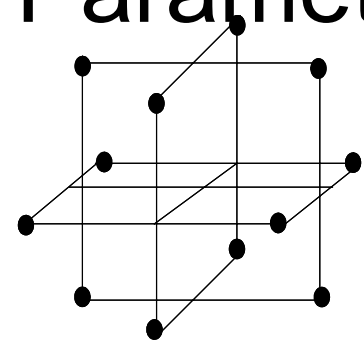
# RSM Pareto-Front GP Parameters- Step 4



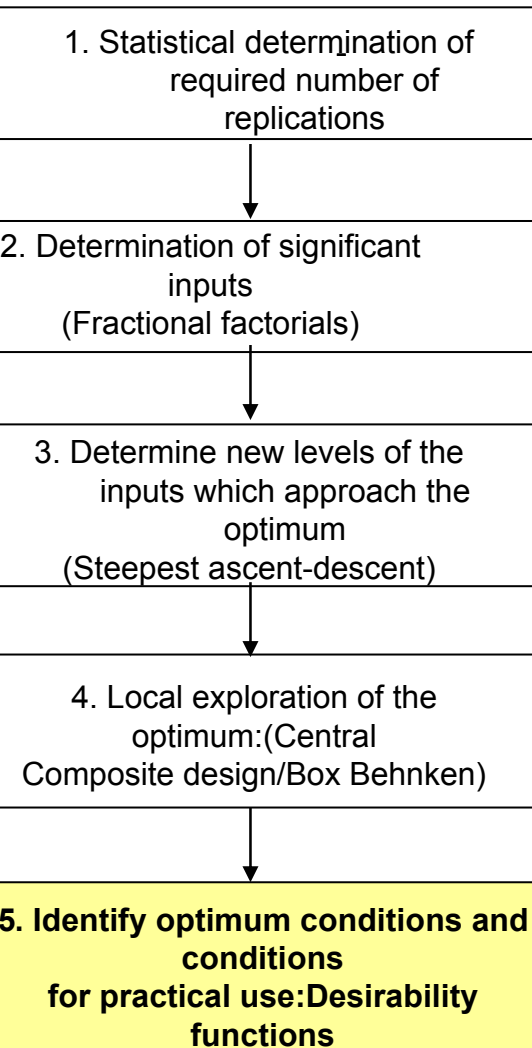
| Factor                                      | Estmate | Prob> t  |
|---|---------|----------|
| Intercept                                   | 5.55    | 1.53E-15 |
| Number of Cascades                          | -0.20   | 3.35E-08 |
| Number of Generations                       | -0.19   | 4.93E-08 |
| Population Size                             | -0.23   | 1.75E-08 |
| Number of Cascades*Number of Cascades       | 0.10    | 1.96E-05 |
| Number of Cascades*Number of Generations    | 0.04    | 0.002823 |
| Number of Generations*Number of Generations | 0.12    | 6.95E-06 |
| Number of Cascades*Population Size          | 0.03    | 0.007175 |
| Population Size*Population Size             | 0.12    | 6.28E-06 |

$$y = 5.55 - 0.2 X_1 - 0.19 X_2 - 0.23 X_3 + 0.04 X_1 X_2 + 0.03 X_1 X_3 + 0.1 X_1^2 + 0.12 X_2^2 + 0.12 X_3^2$$

# RSM Pareto-Front GP Parameters- Step 4

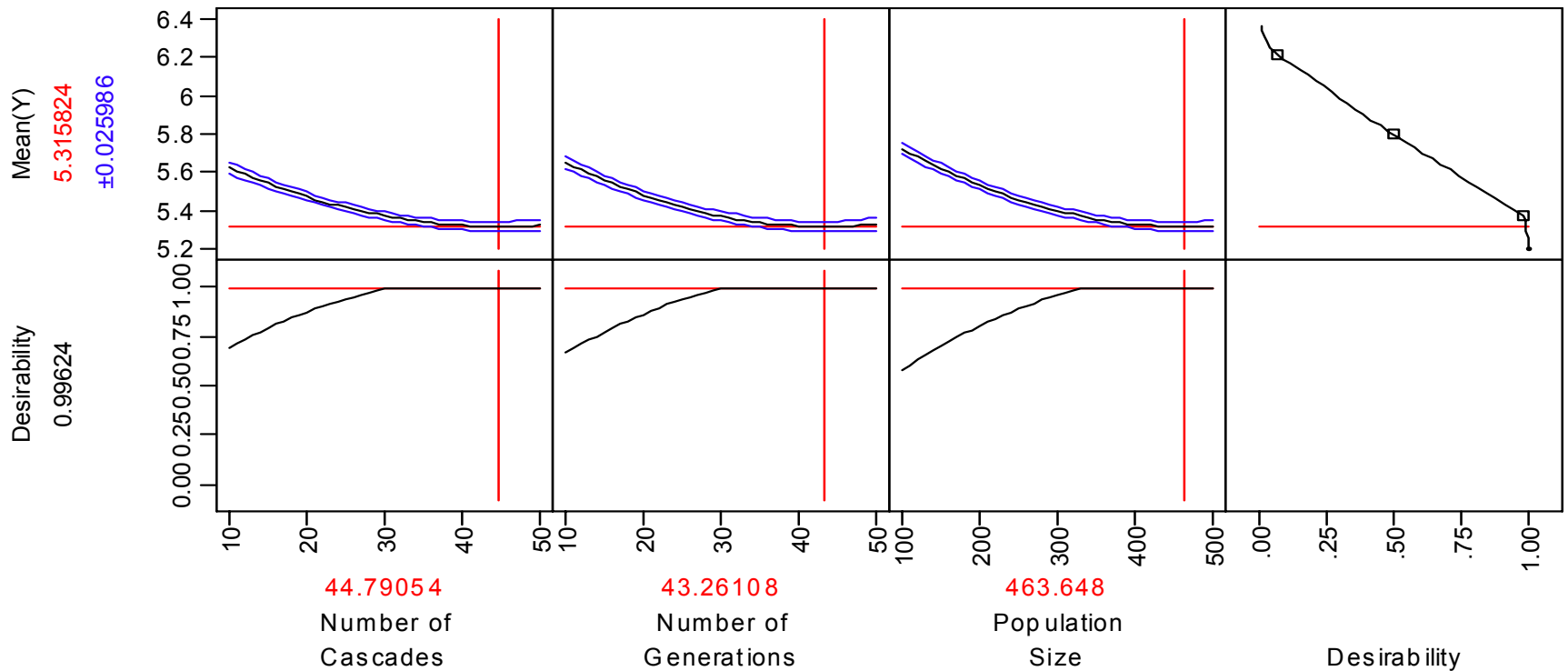


# RSM Pareto-Front GP Parameters- Step 5



The desirability approach : assigns a “score” to a set of responses and chooses factor settings that maximize that score.  
assigns numbers between 0 completely undesirable value and 1

# RSM Pareto-Front GP Parameters- Step 5



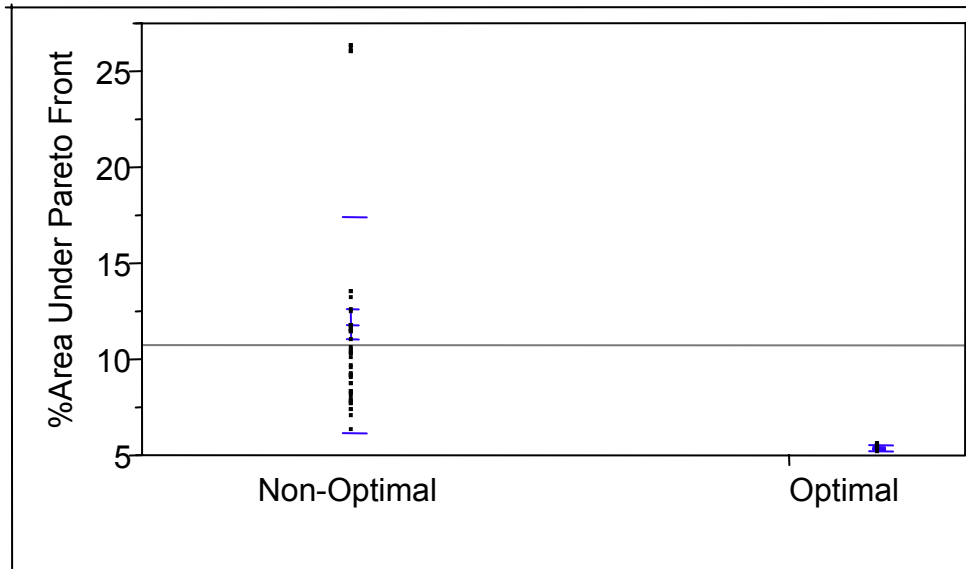
| Factor                                 | Optimal Value |
|--|---------------|
| x1 - Number of cascades                | 45            |
| x2 - Number of generations             | 44            |
| x3 - Population size                   | 464           |
| x4 - Probability of function selection | 0.53          |
| x5 - Size of archive in % of pop. size | 75            |
| Number of replications                 | 10            |

# Compare Optimal - Non-optimal Parameters

(Number of cascades, Number of generations, Population size, Probability of function selection, Archive Size)

**Optimal:** (45, 44, 464, 0.53, 75) 10 replications

**Non-optimal:** (10, 25, 100, 0.6, 75) 50 replications



different medium-sized data set

statistically significant difference  
Welch Anova  
Prob>F < 0.0001

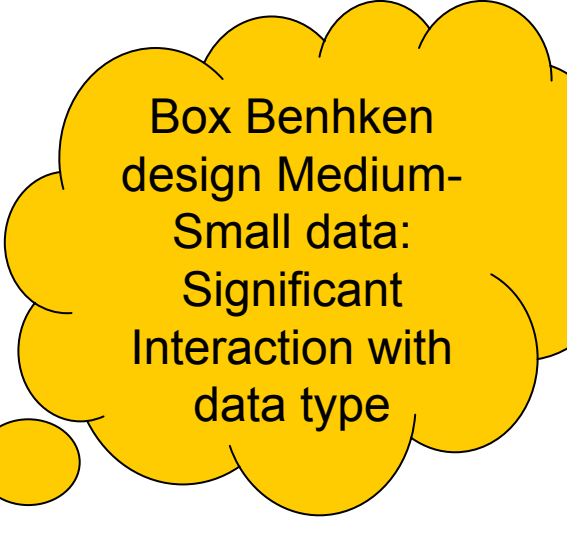
| Level       | Number | Mean  | Std Dev | Lower 95% | Upper 95% |
|-------------|--------|-------|---------|-----------|-----------|
| Non-Optimal | 50     | 11.79 | 5.62    | 10.19     | 13.38     |
| Optimal     | 10     | 5.35  | 0.12    | 5.26      | 5.44      |

# Robustness

Are there GP parameters that are insensitive to the type of data set?

## Example considering small and medium data set

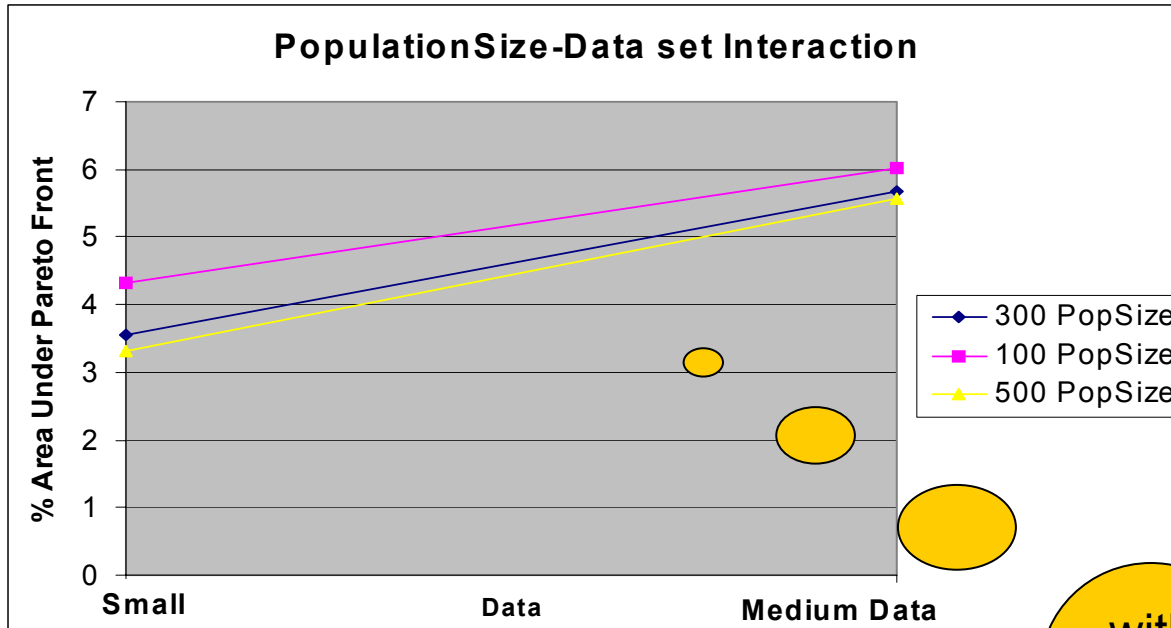
| Term   | Estimate      | Prob> t        |
|--|---------------|----------------|
| Intercept                                    | 4.4440        | 0.00000        |
| Number of Cascades(10,50)                    | -0.2782       | 0.00000        |
| Number of Generations(10,50)                 | -0.2724       | 0.00000        |
| Population Size(100,500)                     | -0.3645       | 0.00000        |
| Number of Cascades*Number of Cascades        | 0.1327        | 0.01751        |
| Number of Cascades*Number Generations        | -0.0063       | 0.89832        |
| Number Generations*Number Generations        | 0.1668        | 0.00426        |
| Number of Cascades*Population Size           | 0.1297        | 0.01594        |
| Number of Generations*Population Size        | -0.0040       | 0.93440        |
| Population Size*Population Size              | 0.2119        | 0.00064        |
| DataSet[Medium]                              | 1.0139        | 0.00000        |
| <b>DataSet[Medium]*Number of Cascades</b>    | <b>0.0757</b> | <b>0.04082</b> |
| <b>DataSet[Medium]*Number,of Generations</b> | <b>0.0825</b> | <b>0.02757</b> |
| <b>DataSet[Medium]*Population Size</b>       | <b>0.1388</b> | <b>0.00088</b> |



Box Benhken  
design Medium-  
Small data:  
Significant  
Interaction with  
data type

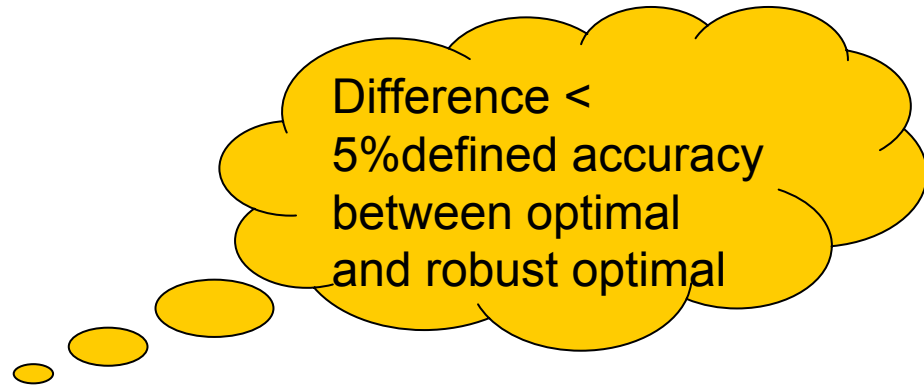
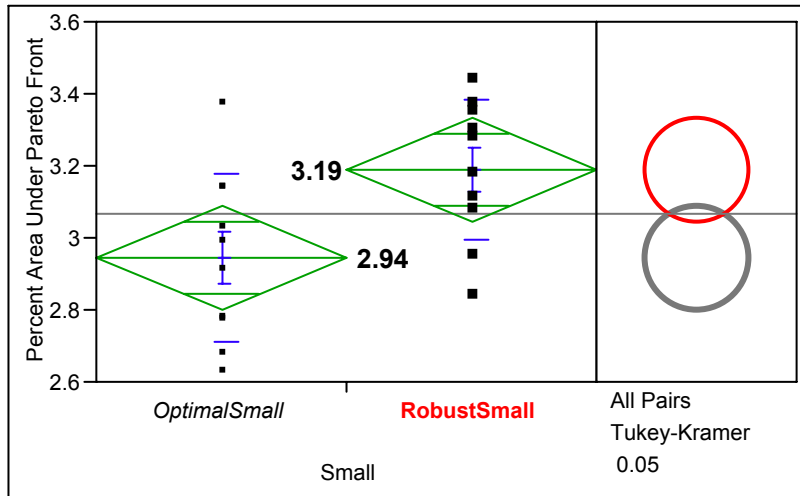
**Possible robustness:**  
**Population size**  
**number of cascades**  
**number of generations**

# Robustness with Respect to Population Size



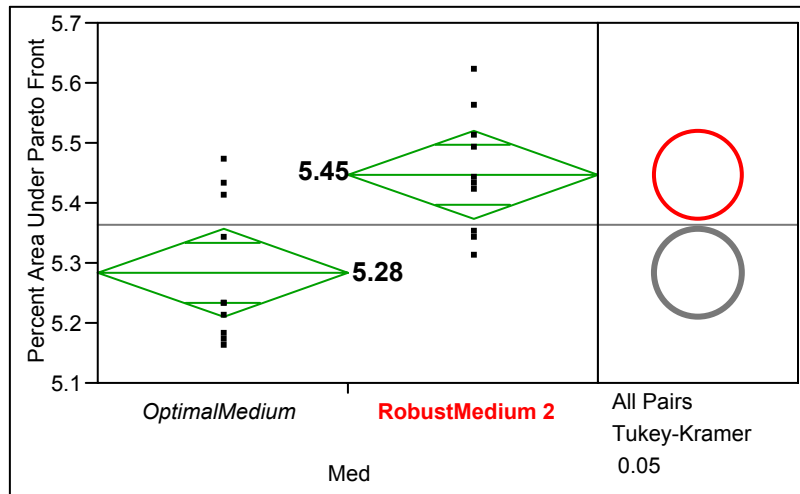
within each type of data set, there is little sensitivity from reducing the population size from 500 to 300.

# Robustness- Population size



**Small Optimal:** (43, 47, **429**, 0.53, 75)

**Small Robust-Pop:**(43, 47, **300**, 0.53, 75)



**Medium Optimal:** (45, 44, **464**, 0.53, 75)

**Medium Robust-Pop:**(45, 44, **300**, 0.53, 75)

Reduction of the computation time

- 52% for the small data

- 48% for the medium data

# Robustness of Small, Medium and Large Data Sets

|                           |                           | Optimum   | Robust_PopSize | Robust_all |
|---------------------------|---------------------------|-----------|----------------|------------|
|                           | Small                     |           | 43,47,429      | 43,47,300  |
| Mean%AreaUnderParetoFront |                           | 2.94      | 3.19           | 3.16       |
| %Saving-Computation Time  |                           |           | 52%            | 51%        |
| Medium                    |                           | 45,44,464 | 45,44,300      | 30,30,300  |
|                           | Mean%AreaUnderParetoFront | 5.28      | 5.45           | 5.6        |
|                           | %Saving-Computation Time  |           | 48%            | 55%        |
| Large                     |                           | 50,50,490 | 50,50,300      | 40,40,400  |
|                           | Mean%AreaUnderParetoFront | 11.47     | 11.85          | 12.05      |
|                           | %Saving-Computation Time  |           | 35%            | 39%        |

Difference < 5% defined accuracy between optimal and robust optimal

# Conclusion

- Statistical techniques for **optimal GP parameter selection** had been illustrated for symbolic regression, generated by Pareto Front GP
- **Robust GP parameters** have been found for symbolic regression problems based on typical industrial data sets. As a result, the **computational time is significantly reduced** while maintaining high-quality solutions
- The **statistically significant** factors (3 out of 5) are: **number of cascades, number of generations, and population size** (particularly important for robustness)
- The specific robust parameters are **limited** to:
  - **Pareto Front GP** algorithm based on two objectives: accuracy and complexity
  - **Symbolic regression** applications
  - Typical industrial data sets with sizes up to **30 variables and 5000 data points**
  - Generated empirical models with  **$R^2 > 0.8$**

# Acknowledgements

- Arthur Kordon
- Guido Smits
- Jeff Sweeney
- Wayne Zirk