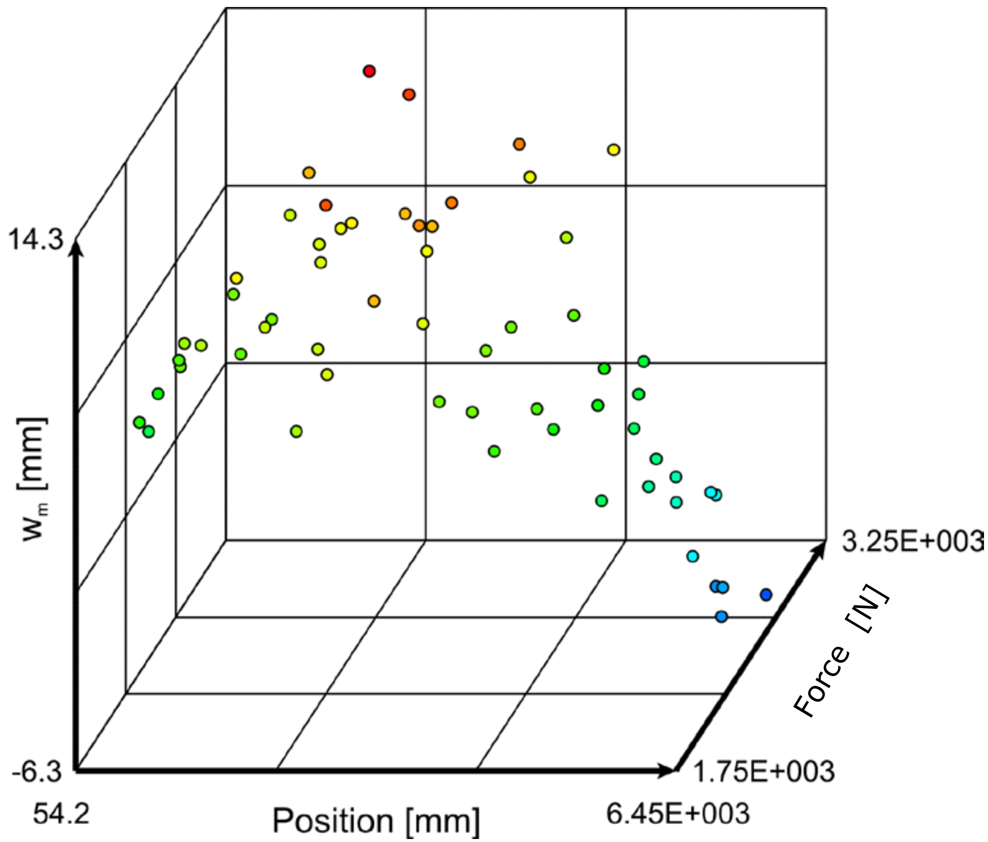
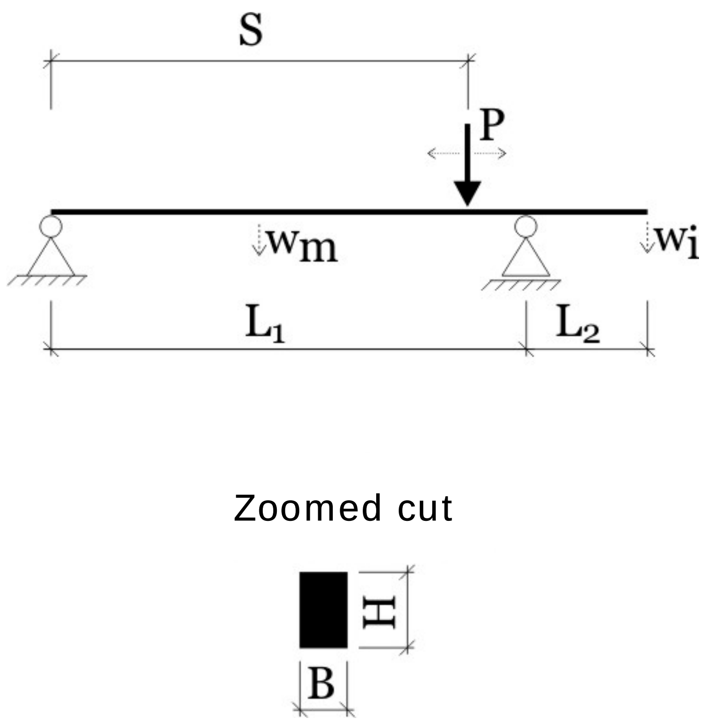
Outline

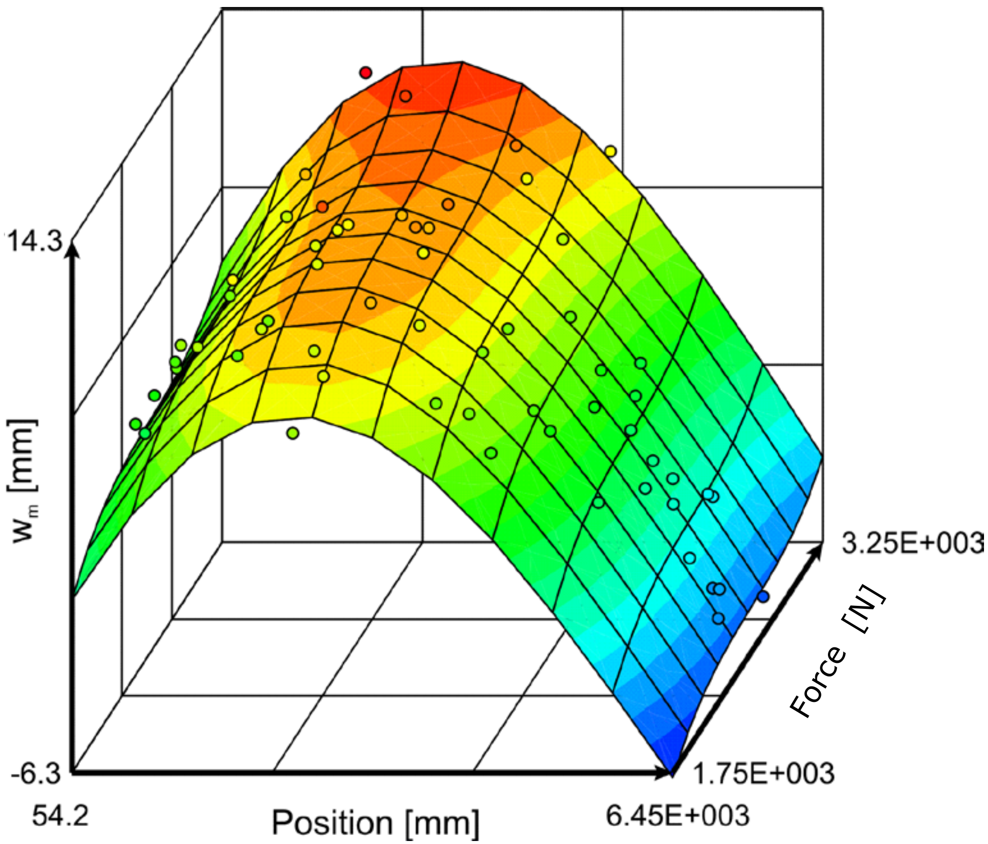
- Introduction
- Part 1: Basics of Statistics
- **Part 2: Regression**
- Part 3: Probabilistic System Analysis using Monte Carlo Methods

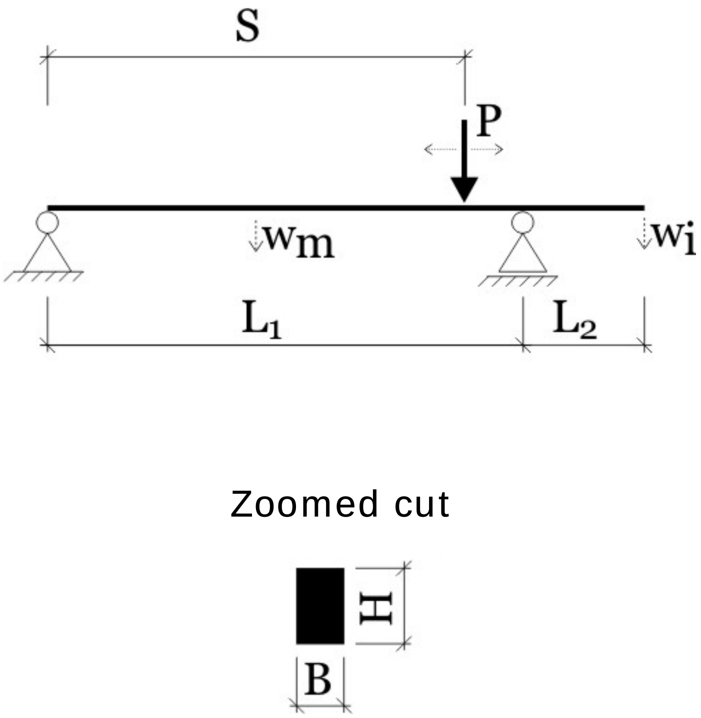
What is a Regression?





What is a Regression?





What is a Regression?

- Synonymic names:
  - meta model
  - response surface
  - surrogate (model)
- Regression is a mathematical model e.g. polynomial equation to fit the deterministic data
- Example – 2nd order polynomial for 2 inputs

$$\tilde{y}_i = c_0 + c_1 \cdot b_{1,i} + c_2 \cdot b_{2,i} + c_3 \cdot b_{1,i}^2 + c_4 \cdot b_{2,i}^2$$

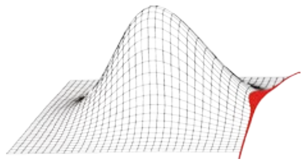
What is a Regression?

- Synonymic names:
  - meta model
  - response surface
  - surrogate (model)
- Regression is a mathematical model e.g. polynomial equation to fit the deterministic data
- Example – 2nd order polynomial for 2 inputs

$$\tilde{y}_i = c_0 + c_1 \cdot b_{1,i} + c_2 \cdot b_{2,i} + c_3 \cdot b_{1,i}^2 + c_4 \cdot b_{2,i}^2$$

○ result

What is a Regression?



- Synonymic names:
  - meta model
  - response surface
  - surrogate (model)
- Regression is a mathematical model e.g. polynomial equation to fit the deterministic data
- Example – 2nd order polynomial for 2 inputs

$$\tilde{y}_i = c_0 + c_1 \cdot b_{1,i} + c_2 \cdot b_{2,i} + c_3 \cdot b_{1,i}^2 + c_4 \cdot b_{2,i}^2$$

○ result

○ input

What is a Regression?

- Synonymic names:
  - meta model
  - response surface
  - surrogate (model)
- Regression is a mathematical model e.g. polynomial equation to fit the deterministic data
- Example – 2nd order polynomial for 2 inputs

$$\tilde{y}_i = c_0 + c_1 \cdot b_{1,i} + c_2 \cdot b_{2,i} + c_3 \cdot b_{1,i}^2 + c_4 \cdot b_{2,i}^2$$

result

coefficients

input

What is a Regression?

- Synonymic names:
  - meta model
  - response surface
  - surrogate (model)
- Regression is a mathematical model e.g. polynomial equation to fit the deterministic data
- Example – 2nd order polynomial for 2 inputs

$$\tilde{y}_i = c_0 + c_1 \cdot b_{1,i} + c_2 \cdot b_{2,i} + c_3 \cdot b_{1,i}^2 + c_4 \cdot b_{2,i}^2 + \underbrace{c_5 \cdot b_{1,i} \cdot b_{2,i}}_{\text{mixed terms}}$$

result ($\tilde{y}_i$)

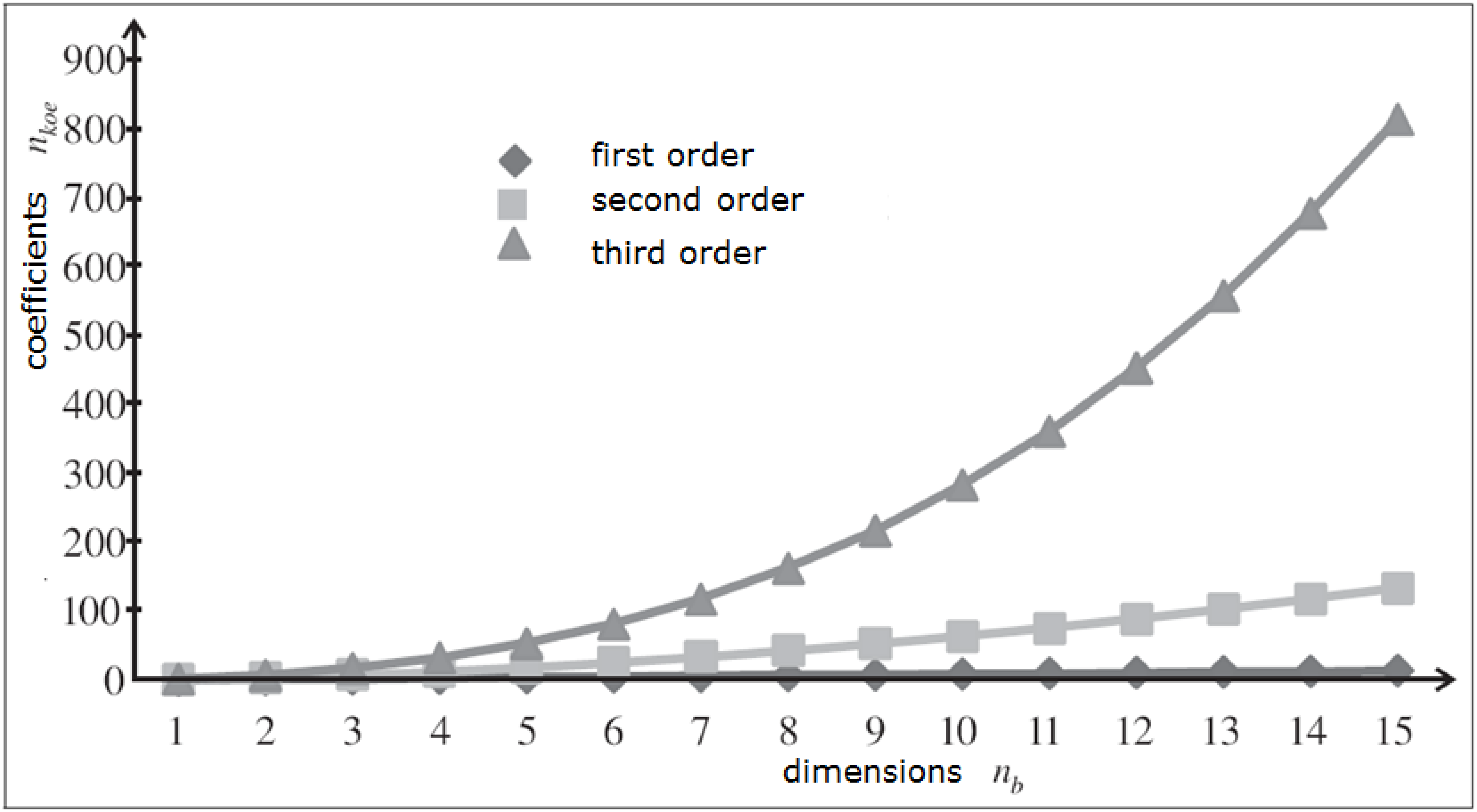
input ($b_{1,i}$, $b_{2,i}$, $b_{1,i}^2$, $b_{2,i}^2$)

coefficients (c_0 … c_4)

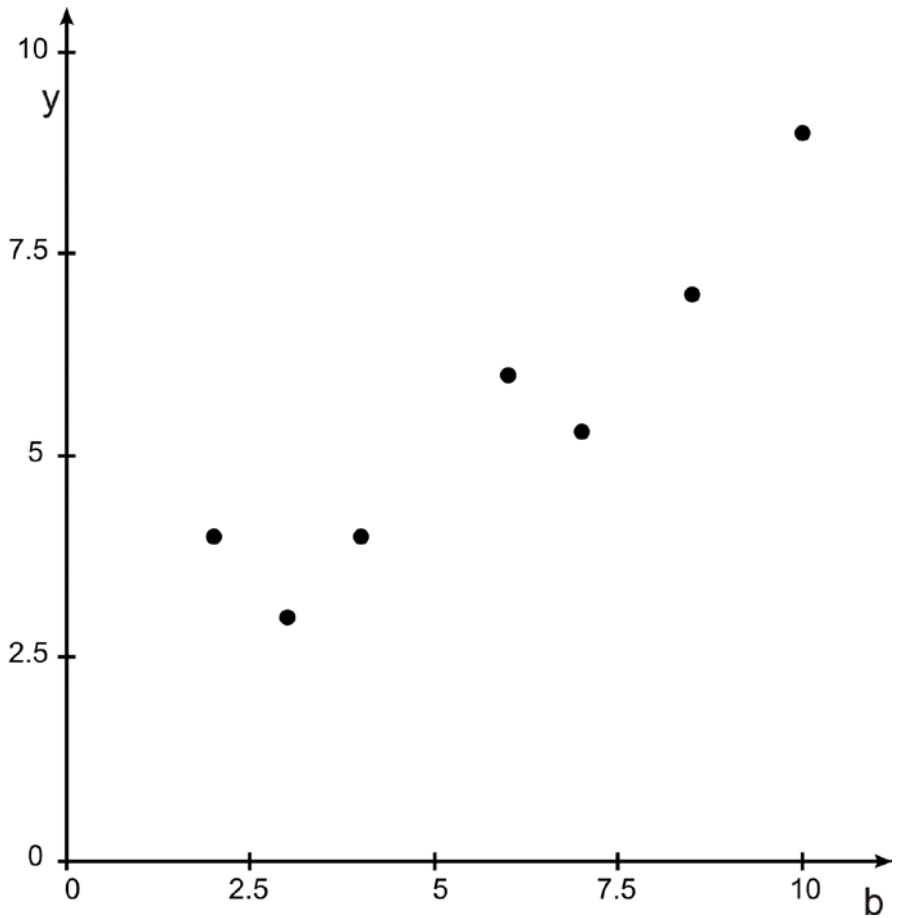
mixed terms

Number of regression coefficients

- Number of data points > number of coefficients !

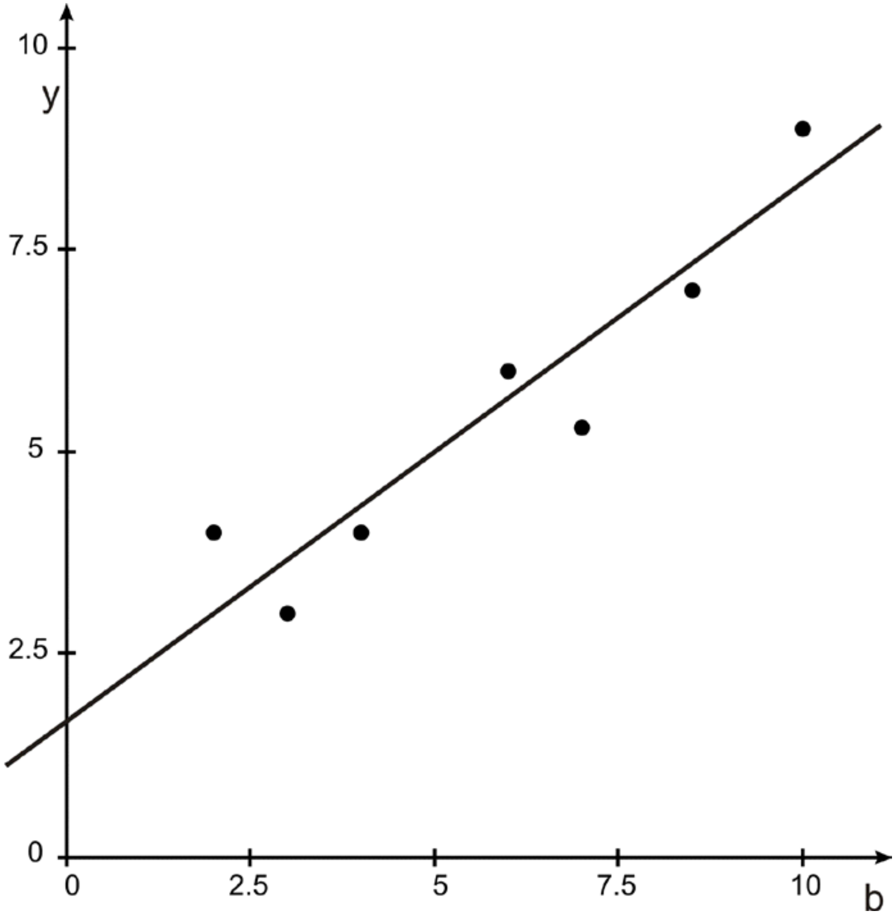


Determine coefficients



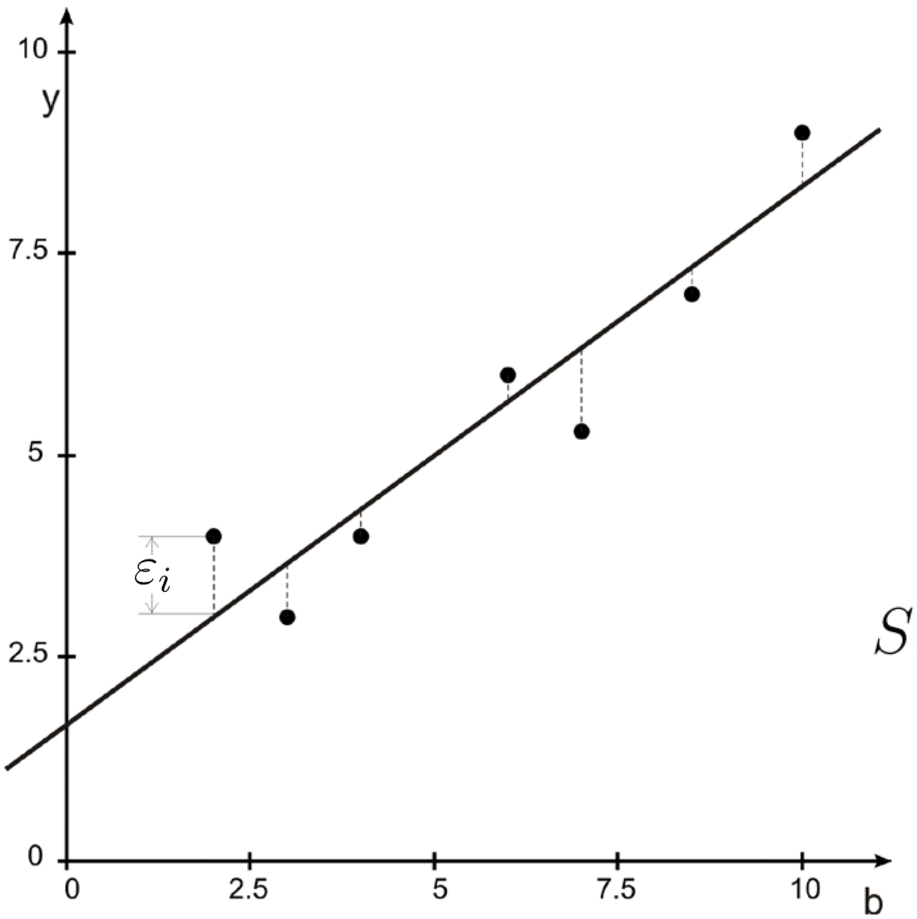
- MCS → point cloud

Determine coefficients



- MCS → point cloud
- Define type of surrogate

Determine coefficients



- MCS → point cloud
- Define type of surrogate
- Compute coefficients to minimize error sum of squares

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Determine fitting quality

- Quality of the fitting between deterministic and surrogate model can be determined by the coefficient of determination

$$R^2 = CoD$$

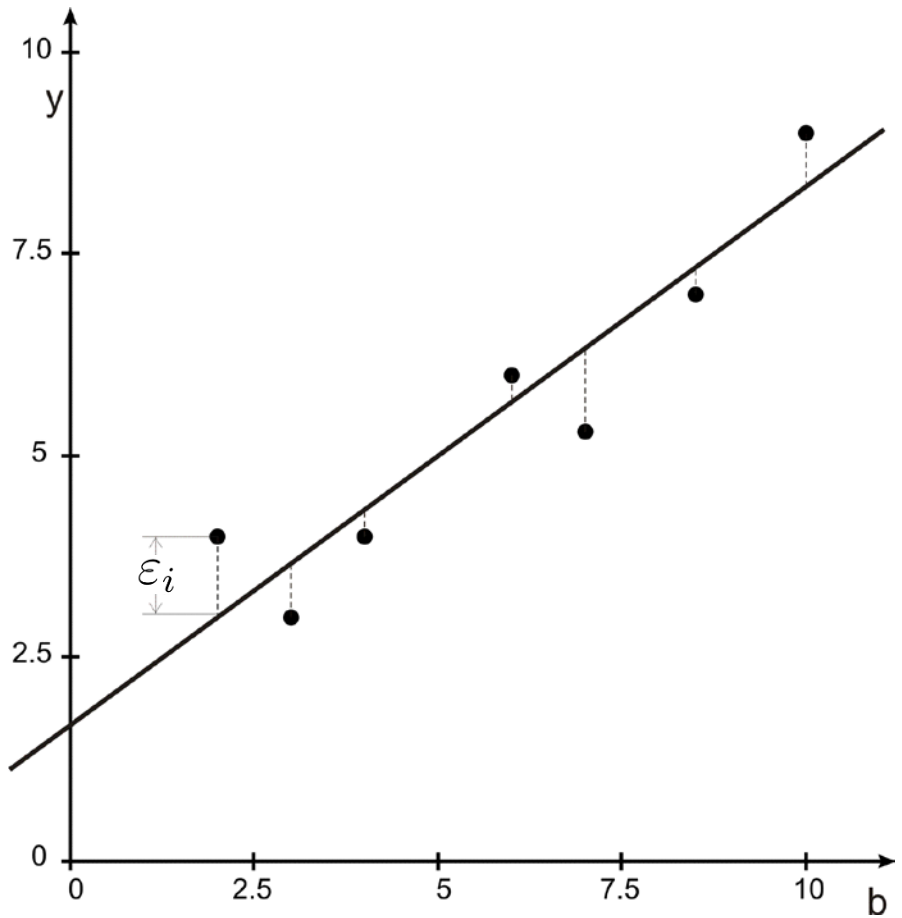
- The CoD describes the part of the total model variation that is included in the regression

$$R^2 = CoD = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- The CoD equals the squared correlation coefficient between the deterministic and the surrogate response

$$R^2 = CoD = r^2(y, \tilde{y})$$

Determine fitting quality



$R^2 = 0.87$

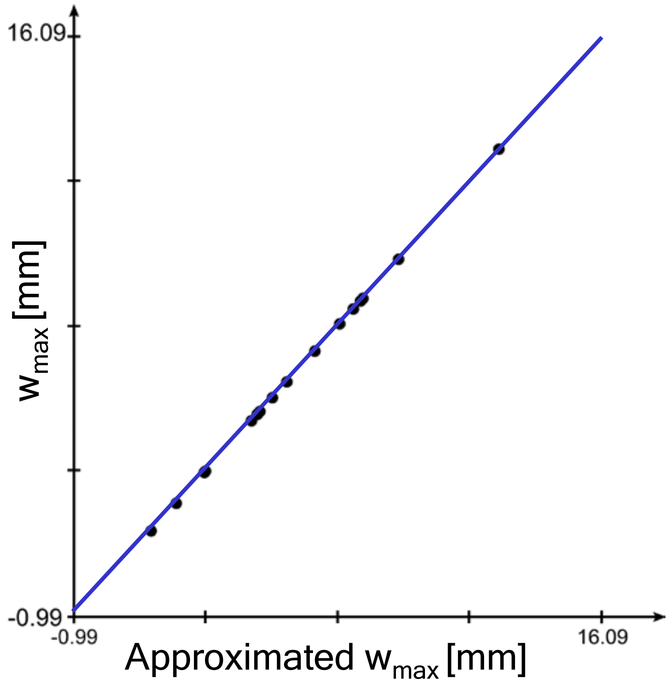
$y = 1,68 + 0,65 * b + \varepsilon$

Determine fitting quality

Example:

- 5 input variables
- 3rd order (w/o mT) → 16 coefficients
- constant input parameter scatter while
- n_{sim} increases

n_{sim}	SCR	R^2
16	1	1.00
30	1.9	0.91
60	3.8	0.86
100	6.3	0.82
200	12.5	0.78
1000	62.5	0.79

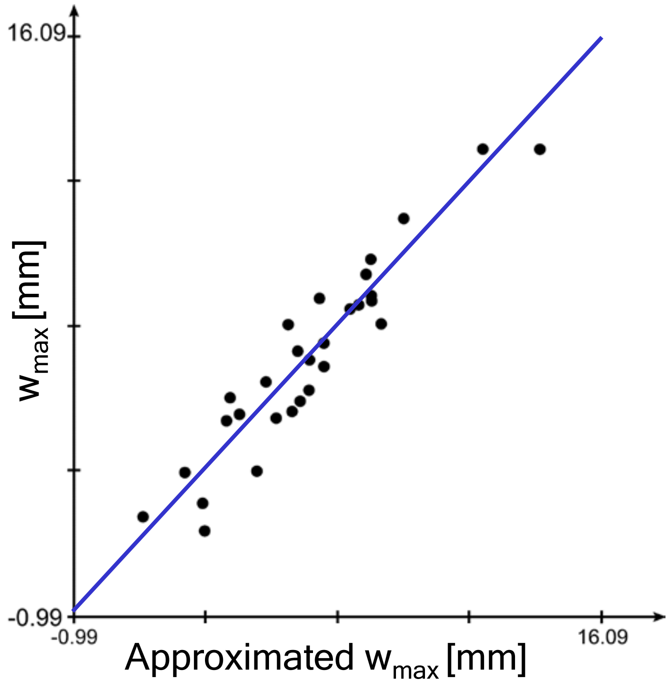


Determine fitting quality

Example:

- 5 input variables
- 3rd order (w/o mT) → 16 coefficients
- constant input parameter scatter while
- n_{sim} increases

n_{sim}	SCR	R^2
16	1	1.00
30	1.9	0.91
60	3.8	0.86
100	6.3	0.82
200	12.5	0.78
1000	62.5	0.79

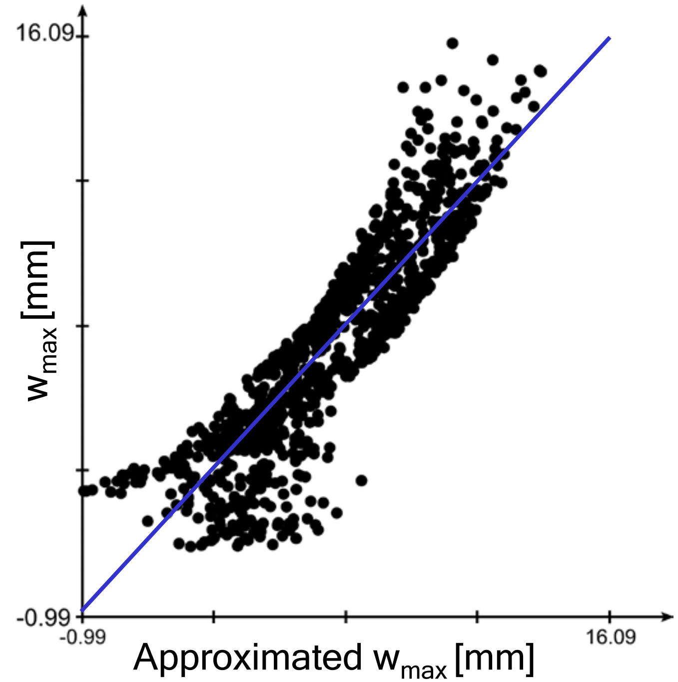


Determine fitting quality

Example:

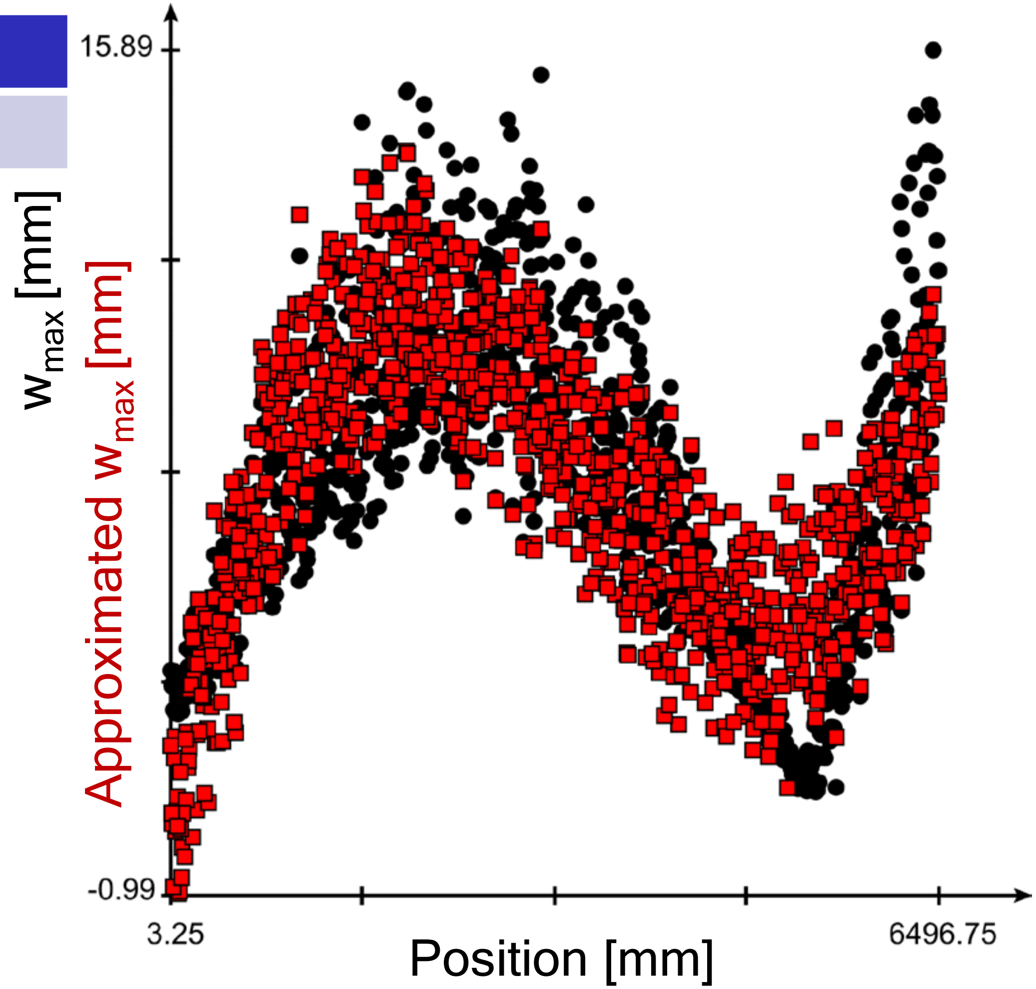
- 5 input variables
- 3rd order (w/o mT) → 16 coefficients
- constant input parameter scatter while
- n_{sim} increases

n_{sim}	SCR	R^2
16	1	1.00
30	1.9	0.91
60	3.8	0.86
100	6.3	0.82
200	12.5	0.78
1000	62.5	0.79



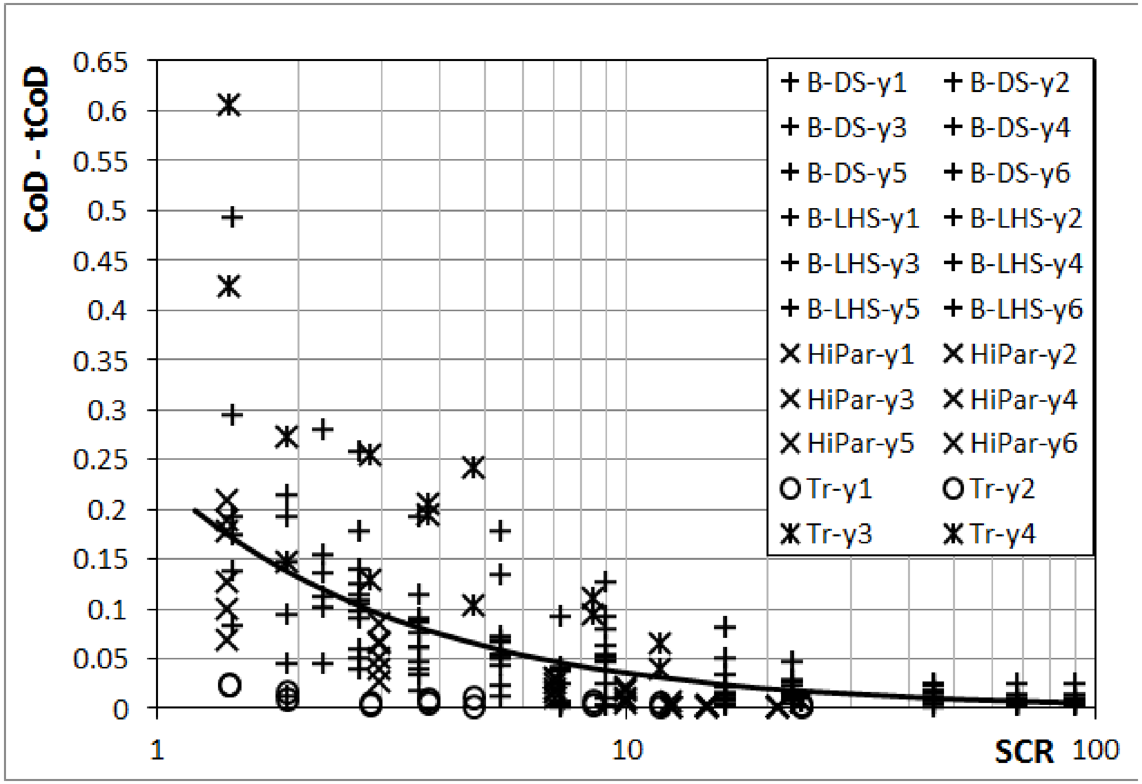
Determine fitting quality

n_{sim}	SCR	R^2
1000	62.5	0.79



CoD as function of SCR (sample to coefficients ratio)

→ Typical saturation behavior of CoD

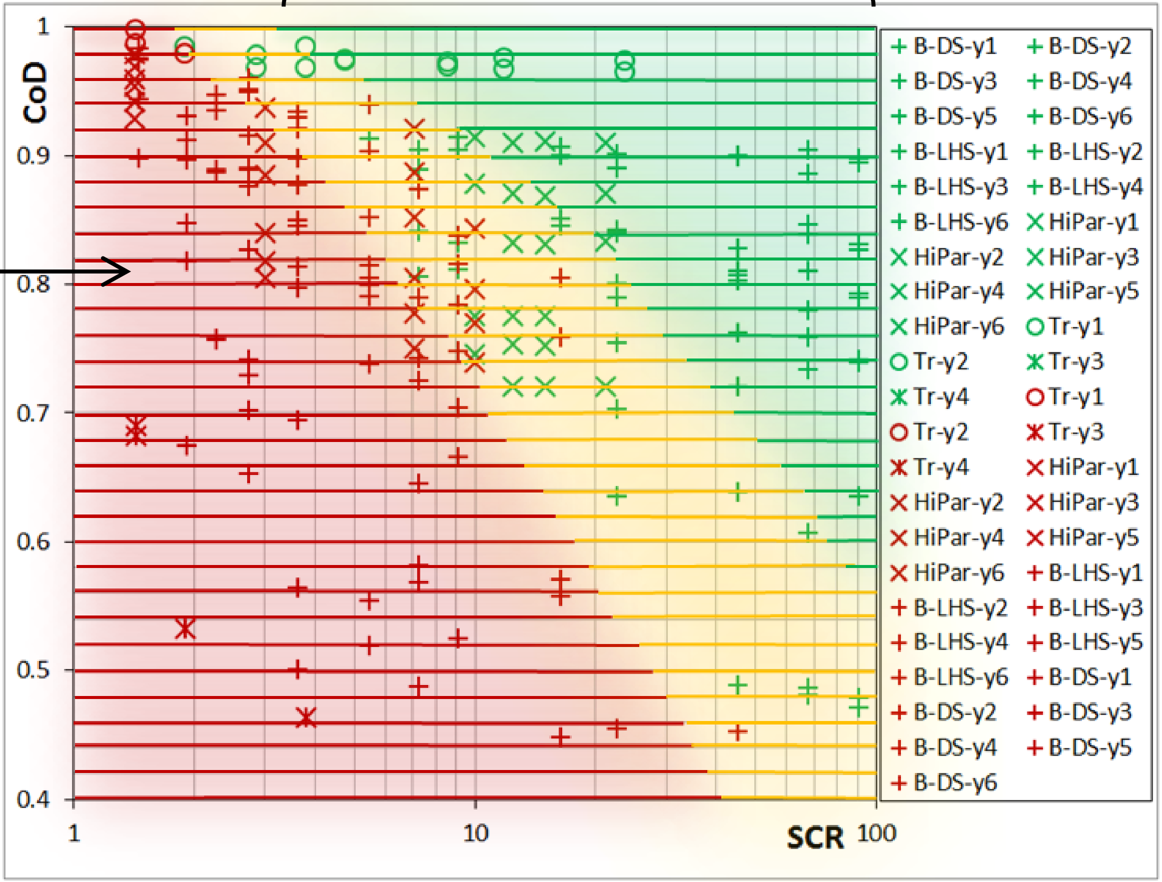


Determine fitting quality

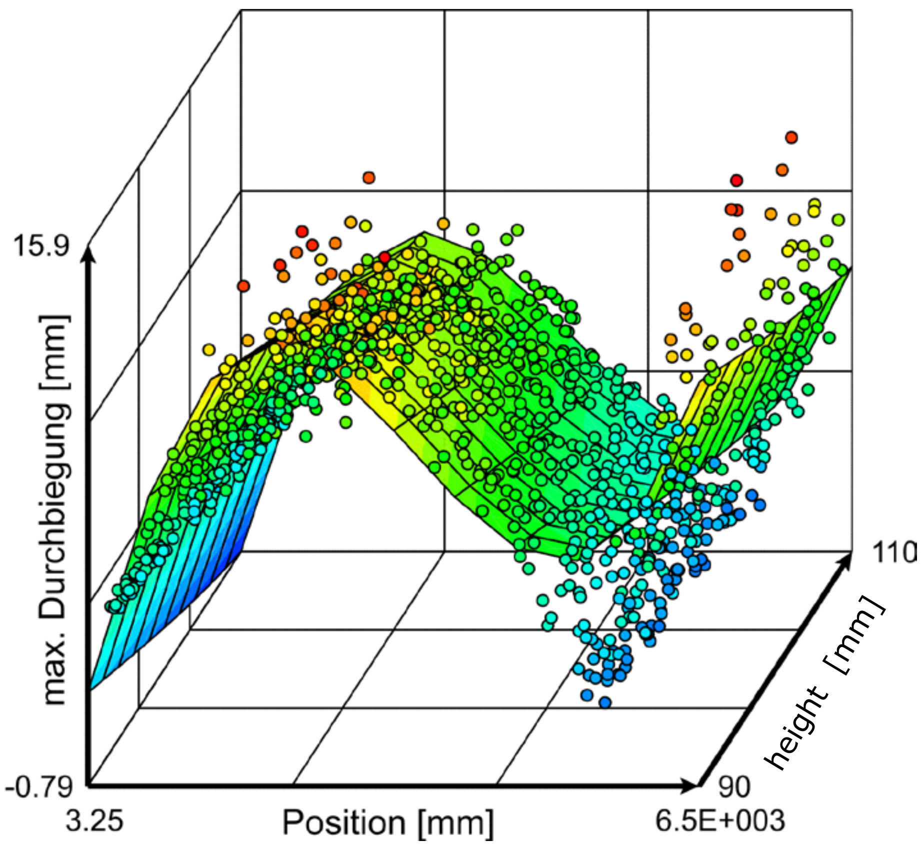
CoD reliability chart

Marginal changes of CoD value by enlarging database

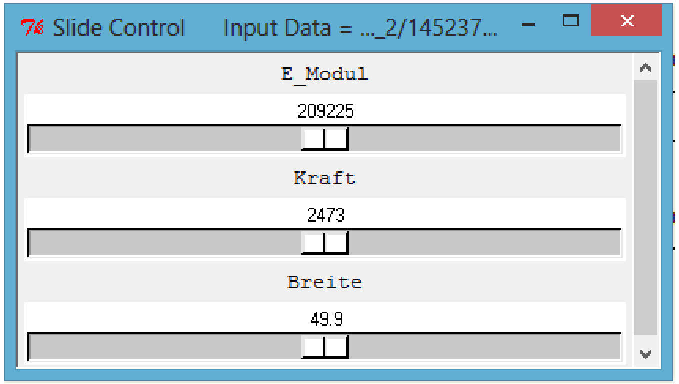
Strong changes of CoD value while enlarging database



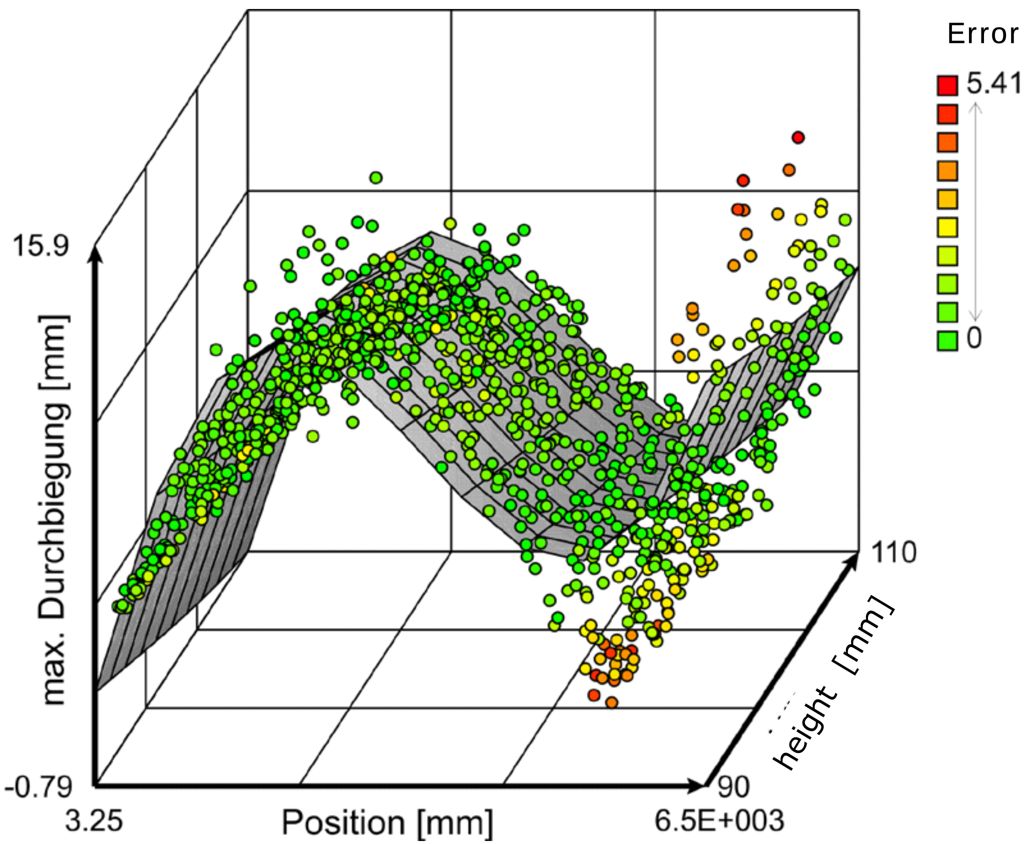
Graphical control



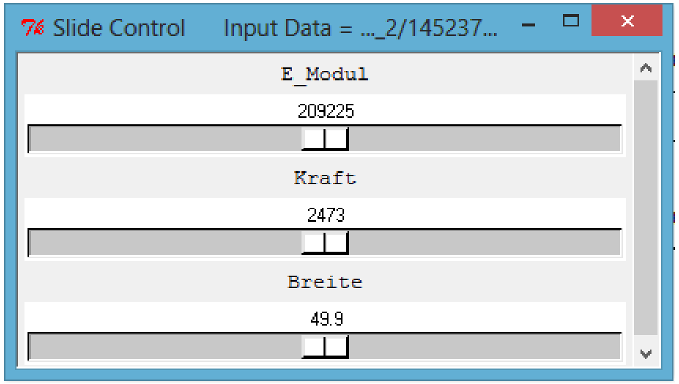
Distance between point and surface does not equal the approximation error since 3 dimensions are not included in the graphic.



Graphical control



Distance between point and surface does not equal the approximation error since 3 dimensions are not included in the graphic.

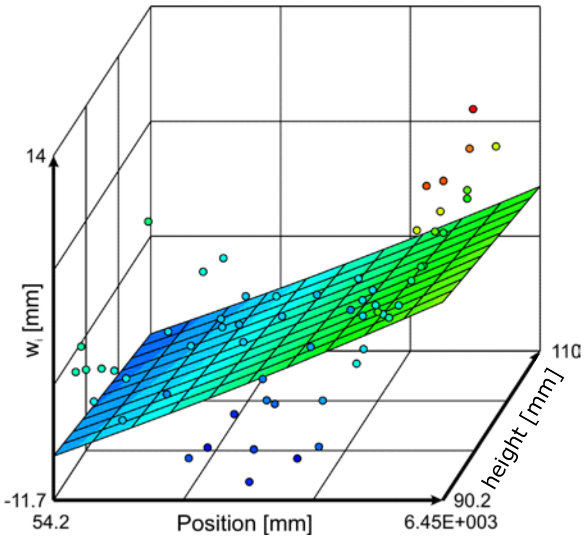


Approximation error is shown as color information

Graphical control

A graphical control is advisable

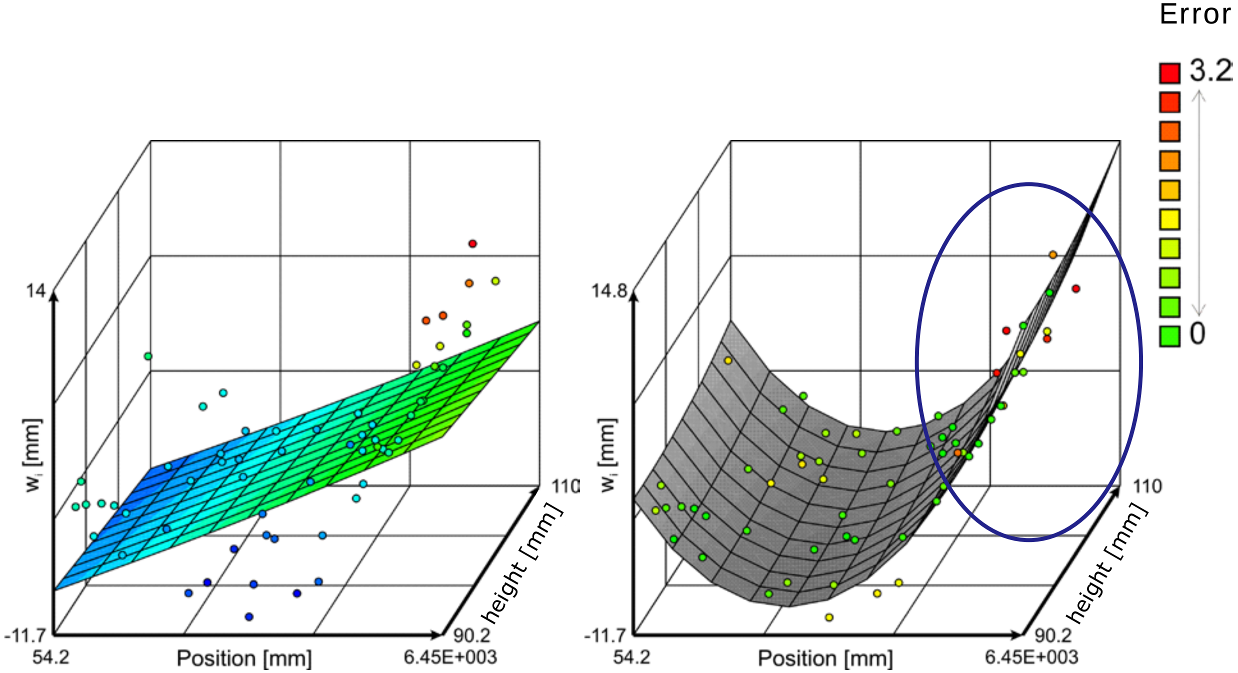
- Does the shape of RS fit to the point cloud?



Graphical control

A graphical control is advisable

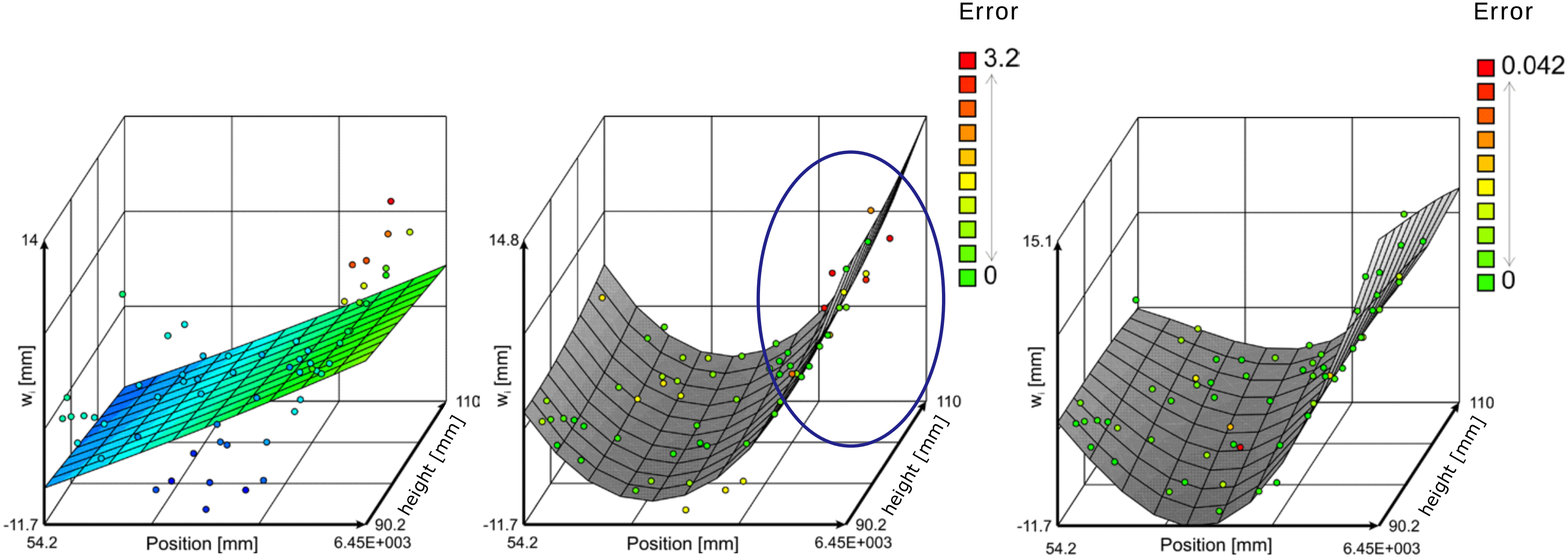
- Does the shape of RS fit to the point cloud?
- Areas with high errors?



Graphical control

A graphical control is advisable

- Does the shape of RS fit to the point cloud?
- Areas with high errors?
- Errors acceptable?



Advantages:

- Widely used criteria
- Comparable between different models (normalized 0...1)
- describes the part of the total model variation that is included in the regression

Disadvantages:

- Value Depends on n
- SCR needs to be high to receive reliable R^2 values
- It does not include information on prediction quality of the model

Applications for Regression

- creation of new data points with the meta model for a fast system analysis
- Replacing the deterministic model with the surrogate if the fit is very good
- Sensitivity analysis with CoI

Work with meta models

When is it possible to work with response surfaces?

High quality of data fitting is necessary

→ (reliable) $R^2 > 0.8$!

→ low approximation error in area of interest

These criteria are no guarantee for a good prediction of new data!

For estimation of prediction quality

→ one needs a validation against new data

→ this data must not be used for model computation

Validation of prediction quality:

No new observations?

→ Split the data set into ***training set*** and a ***validation set***

training set - will be used to compute the response surface

validation set - will be used to validate the prediction error

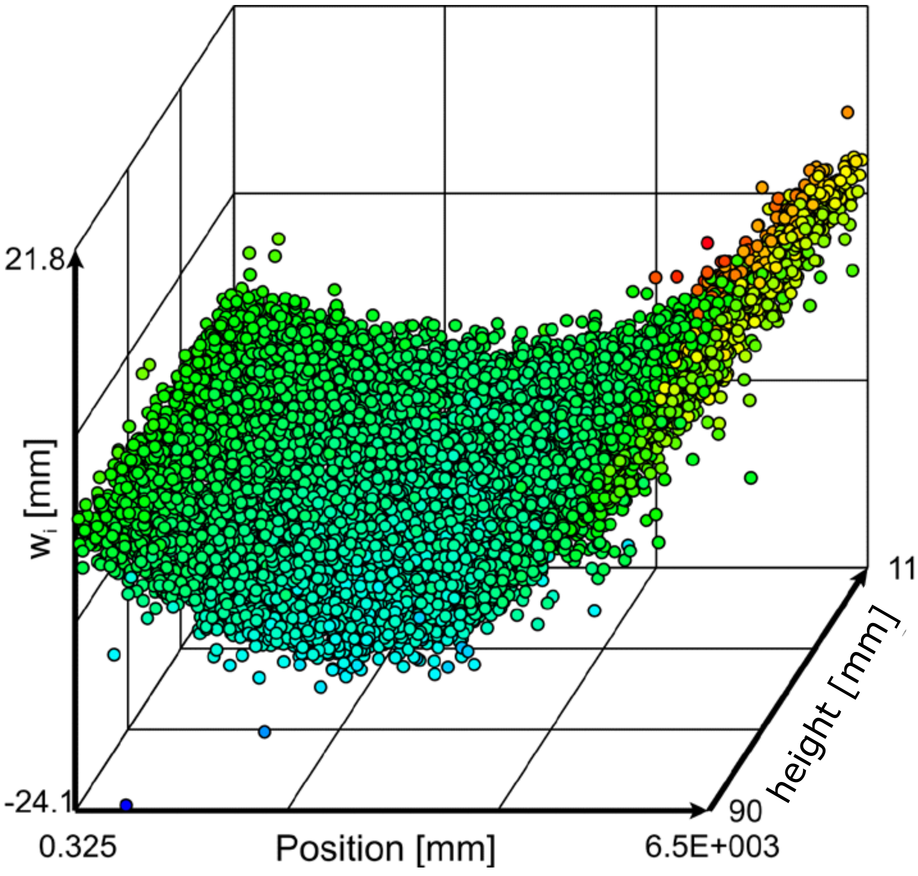
⇩

Cross validation

R^2 prediction or predictive error sum of squares	Different kinds of data splitting

Work with meta models

Validated meta models can be used to compute thousands of new samples in no time.



It is important to make a true validation of the results:

- compare at least one meta model response to the deterministic model response.

CoI – Coefficient of Importance[1]

Evaluates the influence of a single input variable on the result, using the meta model.

$$CoI = R^2 - R_e^2$$

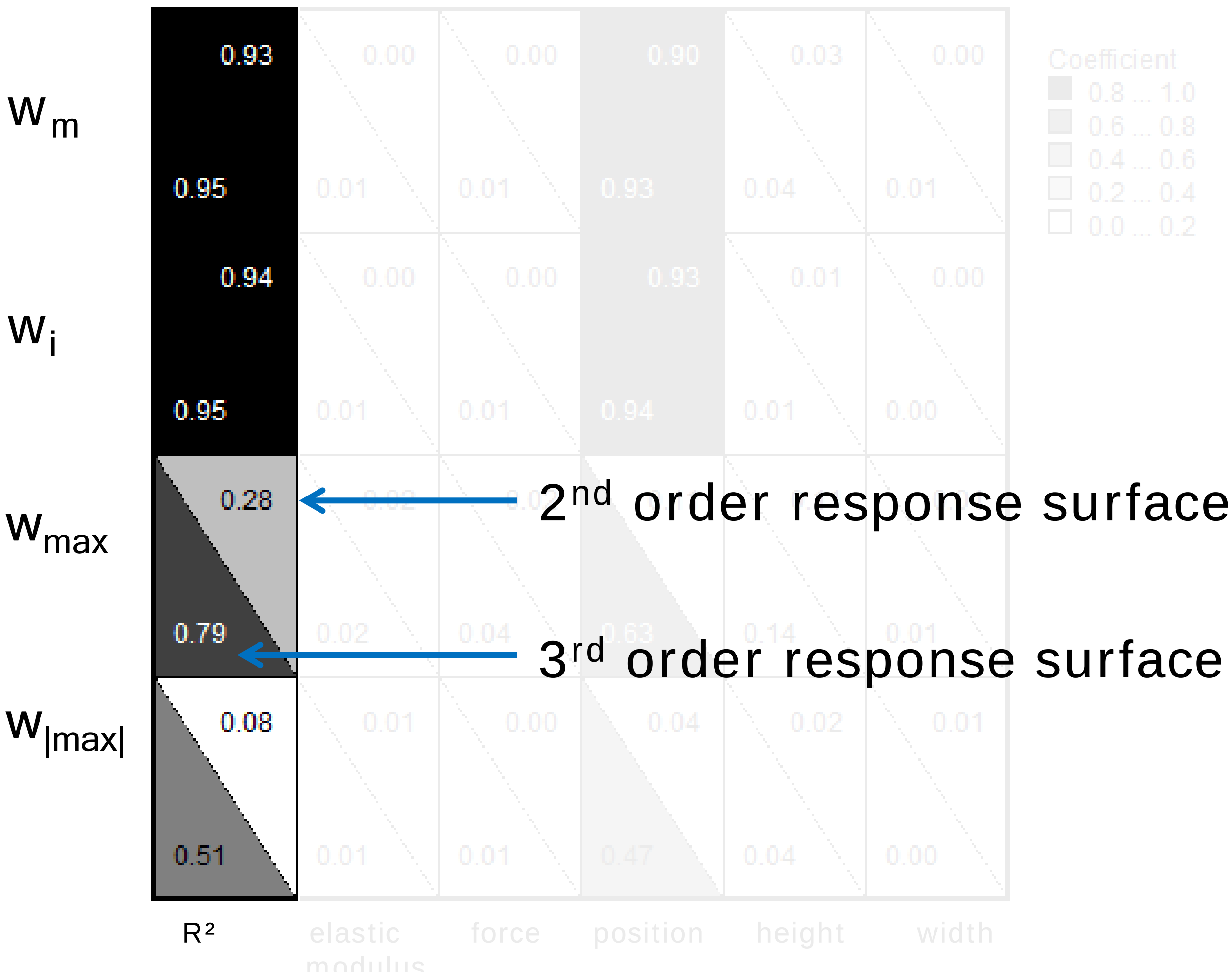
R^2 Coefficient of Determination of the response surface

R_e^2 Coefficient of Determination of a response surface, where the influence of the input variable $\mathbf{b_e}$ is neglected

If the neglected input variable was important, then $R_e^2 << R^2$ and the CoI will be high.

1 WILL, J.; BUCHER, C.: *Statistische Maße für rechnerische Robustheitsbewertungen CAE gestützter Berechnungsmodelle.* Weimarer Optimierungs- und Stochastiktage 3.0, 2006.

Coefficient of Importance



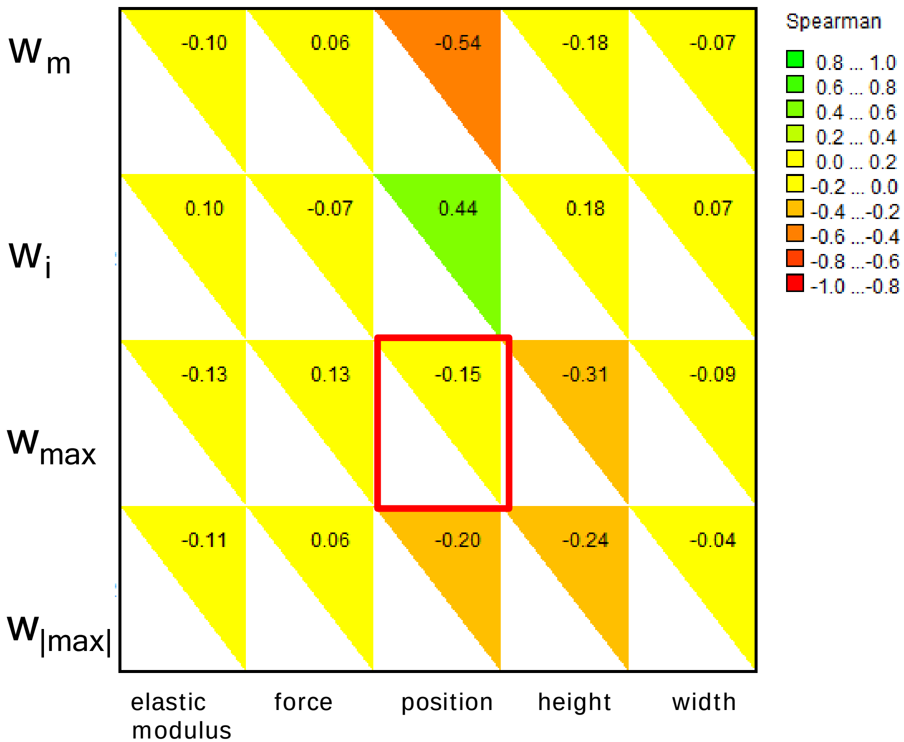
Coefficient of Importance

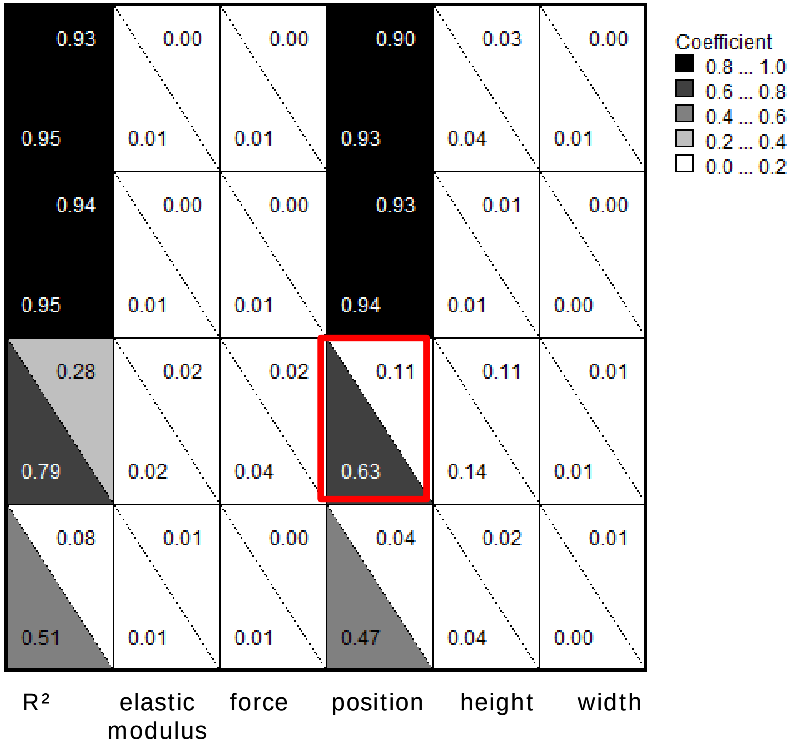
	R^2	elastic modulus	force	position	height	width
W_m	0.93 / 0.95	0.00 / 0.01	0.00 / 0.01	0.90 / 0.93	0.03 / 0.04	0.00 / 0.01
W_i	0.94 / 0.95	0.00 / 0.01	0.00 / 0.01	0.93 / 0.94	0.01 / 0.01	0.00 / 0.00
W_{max}	0.28 / 0.79	0.02 / 0.02	0.02 / 0.04	0.11 / 0.63	0.11 / 0.14	0.01 / 0.01
$W_{\|max\|}$	0.08 / 0.51	0.01 / 0.01	0.00 / 0.01	0.04 / 0.47	0.02 / 0.04	0.01 / 0.00

Coefficient
- 0.8 ... 1.0
- 0.6 ... 0.8
- 0.4 ... 0.6
- 0.2 ... 0.4
- 0.0 ... 0.2

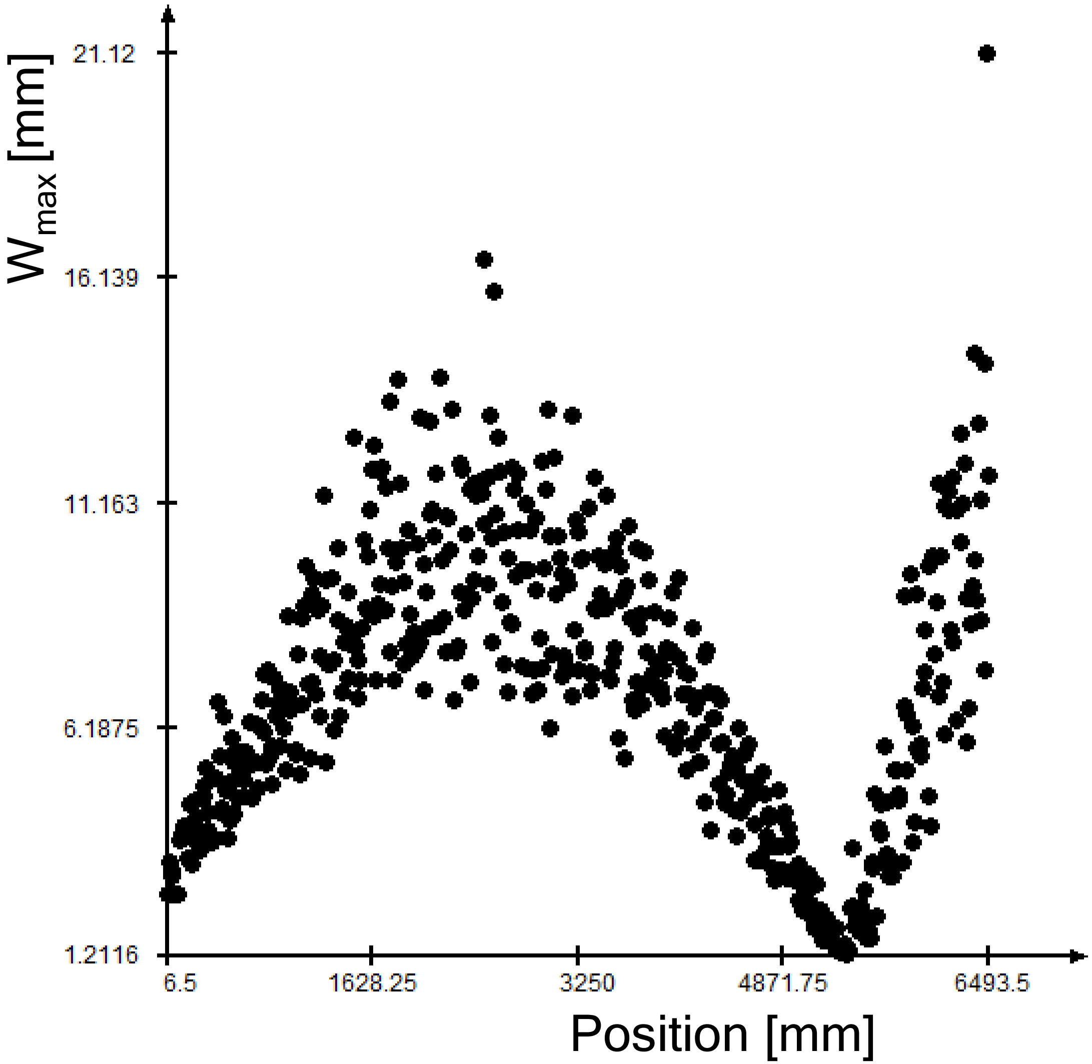
Advantages of CoI compared to correlation coefficient

1. recognition of non monotonic connection





Coefficient of Importance

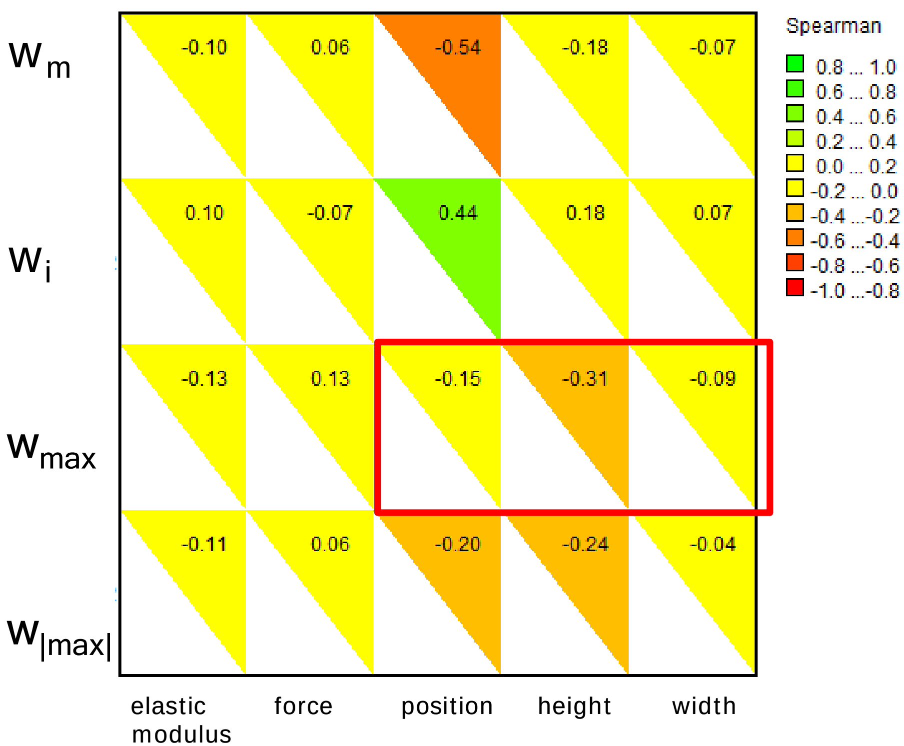


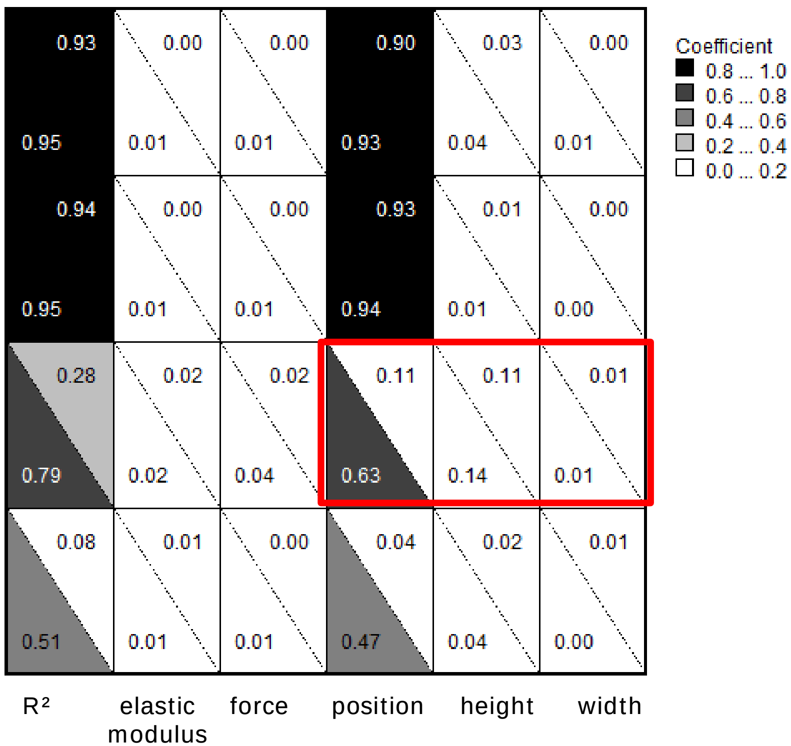
r = -0,15

COI = 0,63 (R^2=0,79)

Advantages of CoI compared to correlation coefficient

1. recognition of non monotonic connection
2. focused sensitivity analysis

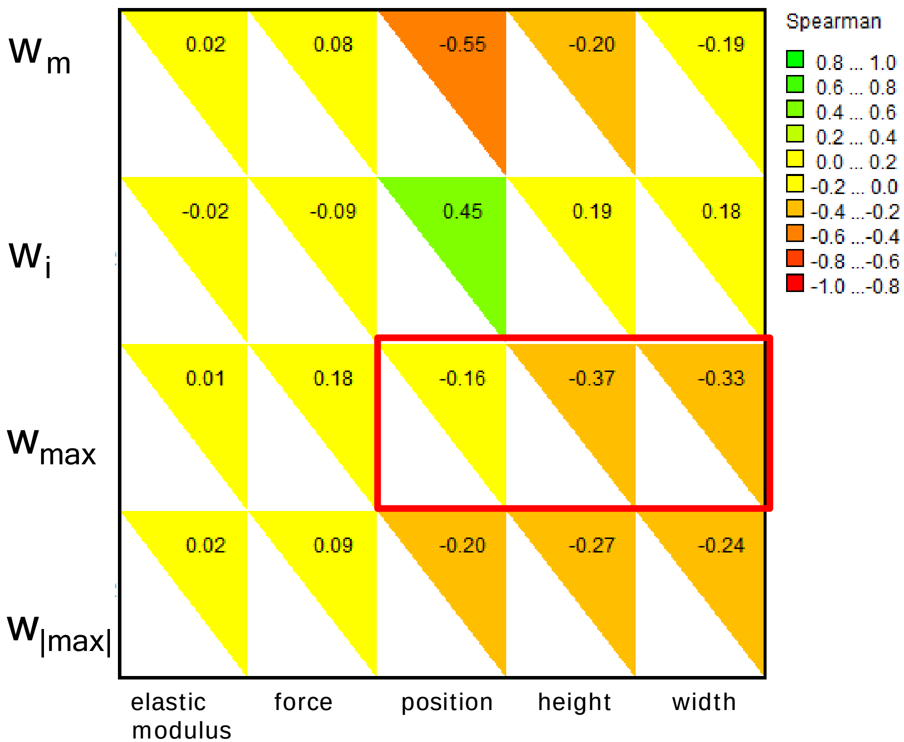


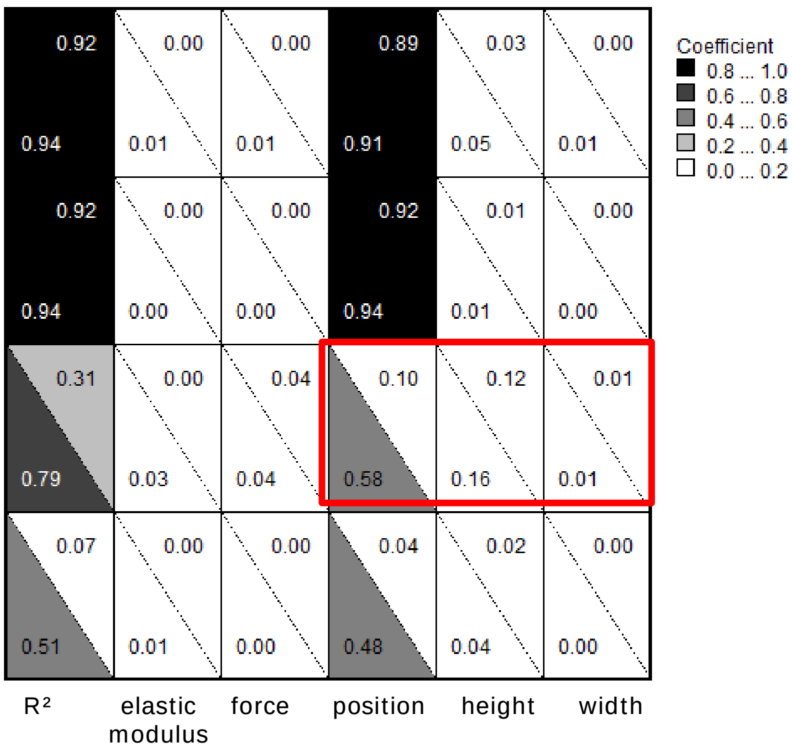


Advantages of CoI compared to correlation coefficient

1. recognition of non monotonic connection

2. focused sensitivity analysis





Box Cox Transformation[1]

$$y = c_o + \sum_{i=1}^{n_b} c_i\, b_i + \sum_{i=1}^{n_b} \sum_{j=1}^{n_b} c_{ij}\, b_i b_j + \varepsilon$$

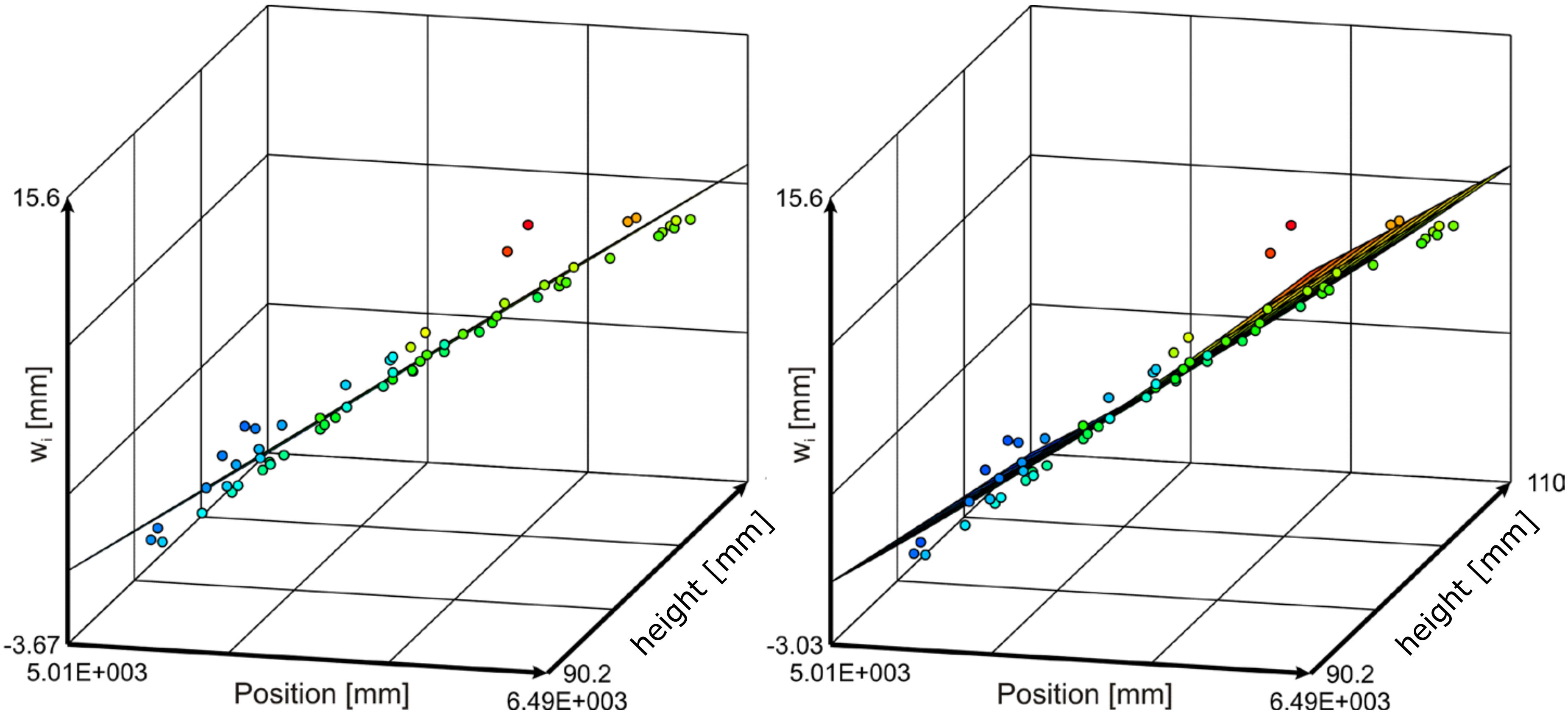
Transformation of ordinate-values to improve approximation

$$y^s = \overline{c_0} + \sum_{k=1}^{n_b} \overline{c_k}\, b_k + \sum_{k=1}^{n_b} \sum_{m=1}^{n_b} \overline{c_{km}}\, b_k b_m + \bar{\varepsilon}$$

Exponent *s* will be selected to further minimize the error sum of squares s=[0.1, ... , 2]

1 Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wassermann, W.: *Applied Linear Statistic Models*. WCB McGraw-Hill, New York, 1996.

Box Cox Transformation



Summary on Regression

Response surfaces are mathematical models to approximate the deterministic system response

Can be used to:

- Check sensitivity with the CoI
- Replace the deterministic model

High model quality is necessary

- $R^2 > 0.8$
- Low error in area of interest
- Use cross validation for a prediction criteria

TECHNISCHE
UNIVERSITÄT
DRESDEN

Faculty of Mechanical Science and Engineering, Institute of Fluid Mechanics, Chair of Turbomachinery and Jet-Propulsion

Tutorial

Introduction into probabilistic methods and their application in engineering sciences with focus on monte carlo and response surface methods

David Pusch
André Beschorner
Robin Schmidt