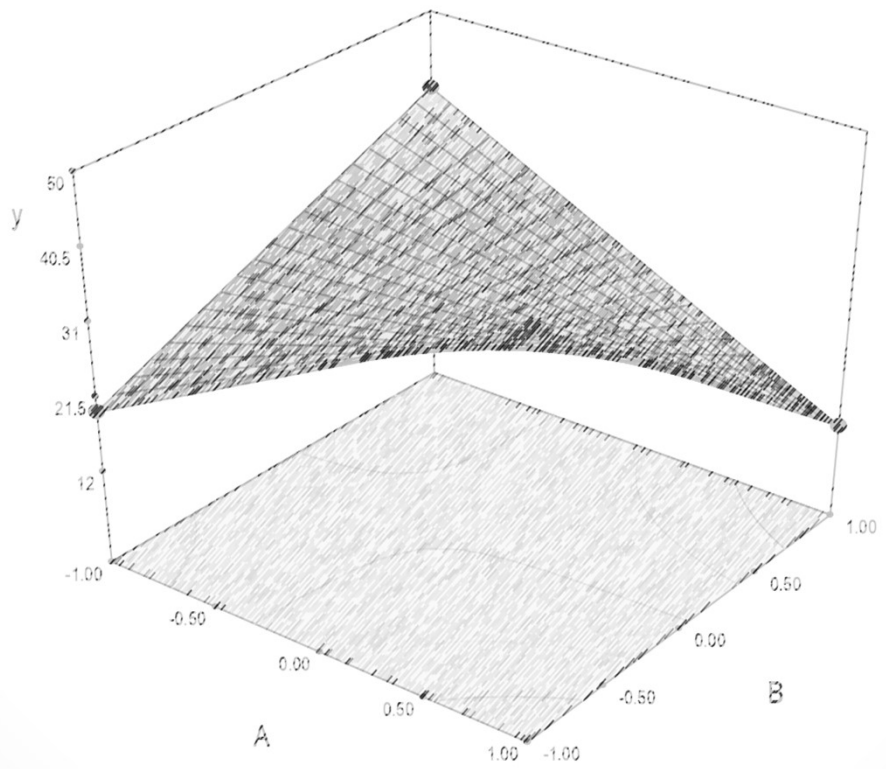
Nowelties
European Joint Doctorate

Design and Analysis of Experiments

Introduction to DOE

Assoc. Prof. Hrvoje Cajner
Department of Industrial Engineering
University of Zagreb
FAMENA
hrvoje.cajner@fsb.hr



Introduction to DOE

Concepts

Basic terms

Models

Analysis

Common models

What is DOE and why you should use it?

- A collection of predetermined process variables setting
- To identify influence of the certain parameters on observed characteristics
- To express the relationship throughout useful model
- To achieve the optimum parameter settings

"When your television set misbehaves, you may discover that a kick in the right place often fixes the problem, at least temporarily. However, for a long term solution the reason for the fault must be discovered. In general, problems can be fixed, or they may be solved. Experimental design catalyzes both fixing and solving."
-- Box, Hunter and Hunter --

Approaches to experimental design

Conventional approach (obsolete)

- Performing experiment without a plan
- Using „one factor at a time” (OFAT) methodology
- Time and resources consuming
- Unable to interpret the interaction between more variables
- Loss of valuable information

Statistical approach (DOE)

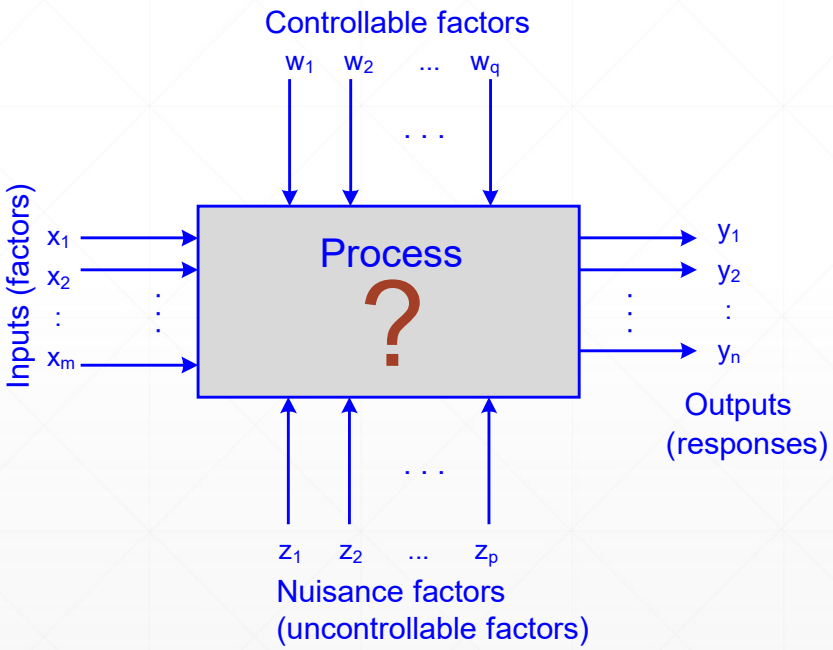
- Applying the DOE methodology – statistical principles
- Reducing the time and resources allocated for the experiments
- Able to interpret the interactions between process variables
- Predicting the values and performing optimization
- Maximizing the information

Basic terms in DOE

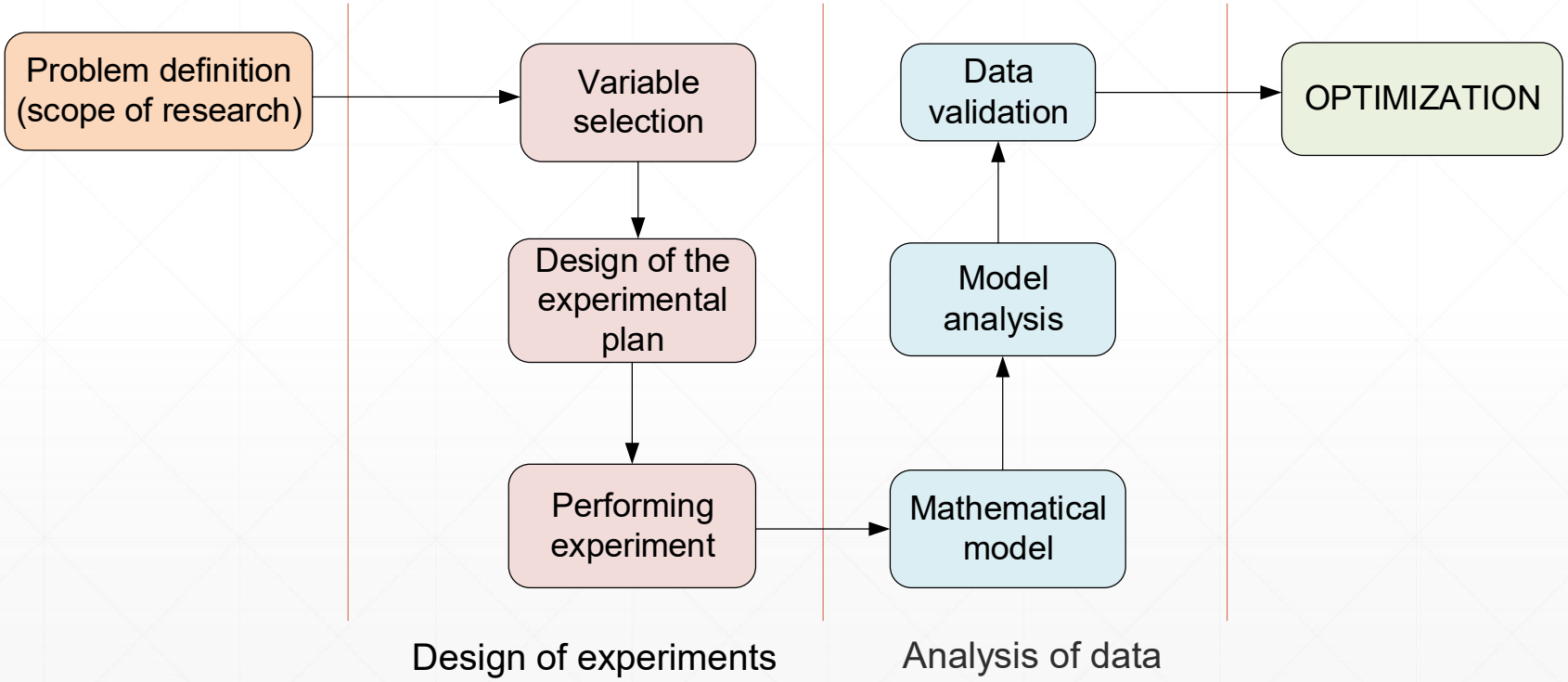
- Factors – independent process variables (IV)
- Level – specific value or settings of factor
- Treatment – factor at specified level
- Response – dependent variable (DV)
- Effect – how much response changes due to factor change
- Interaction – when factors changes simultaneously and have different effect than alone
- Alias – when the estimate of an effect includes the influence of one or more other effects
- Model – mathematical relationship between IV and DV
- Randomization – performing the experimental points in random order
- Blocking – run the experiment in separate blocks whenever the complete randomization isn't possible
- Replication – making multiple experimental runs for each design point

General model of process

- Input variables (factors) – process variables – scope of the research
- Controllable factors – the variables which can be held as constants – not in scope of the research
- Nuisance factors – unavoidable – needs to be minimized
- Output variables – responses which represents the objective – characteristics
- Model: $y_1=f_1(x_1, x_2, \ldots,x_m)$; $y_1=f_2(x_1, x_2, \ldots,x_m)$



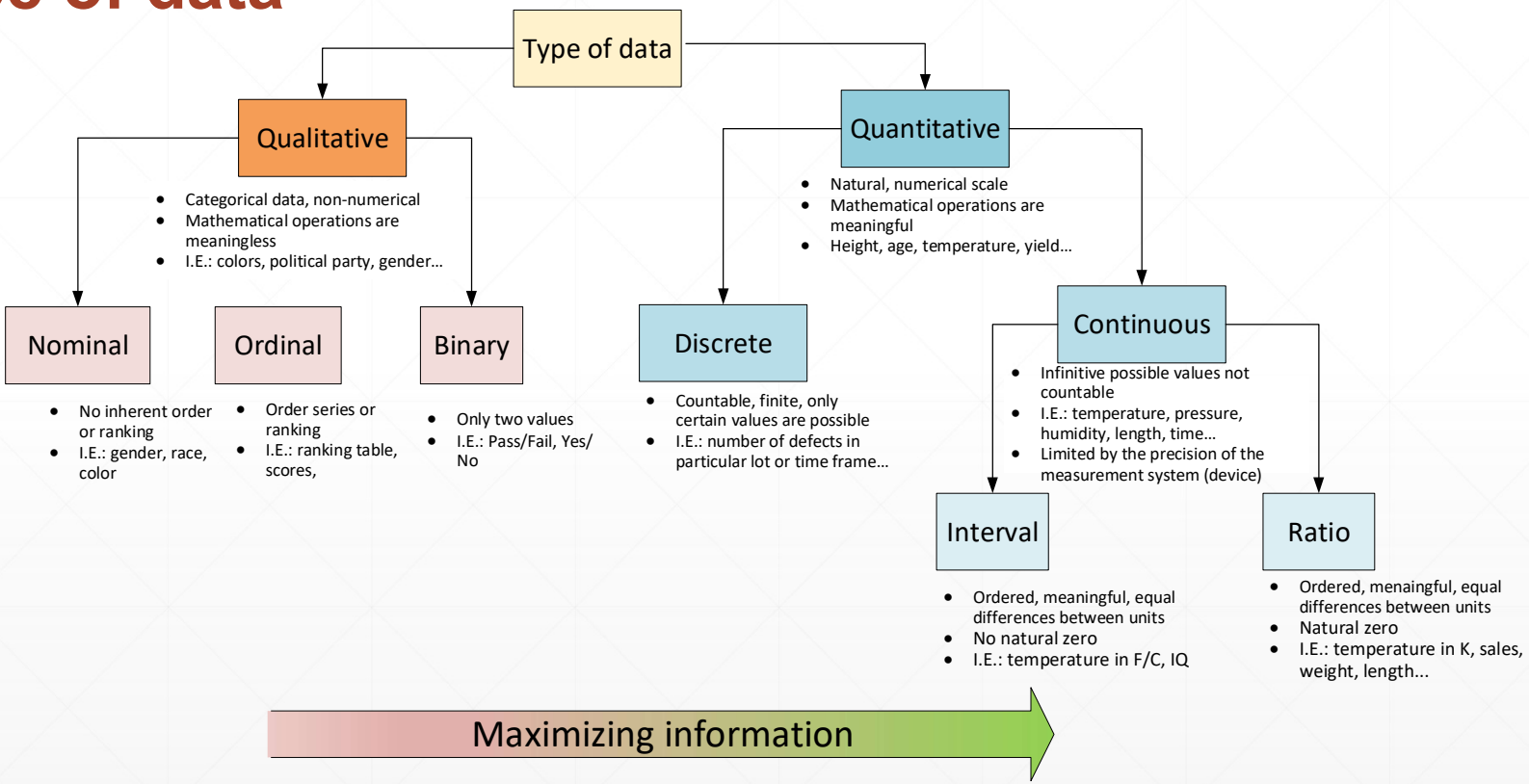
DOE flowchart



Data

- Data should clarify the objective of the research and how it relates to process performance or scope
- It is important to determine:
  - What data should be collected?
  - What type of data is suitable for collecting?
  - How to measure it?
- The data needs to be representative (unbiased)
- The process of sampling (experimenting) should be conducted in random order, so that obtained data isn't contaminated by uncontrollable external influences
- The resolution of the collected data must fit the overall process characteristics

Type of data


Type of data
Qualitative
Categorical data, non-numerical
Mathematical operations are meaningless
I.E.: colors, political party, gender…
Quantitative
Natural, numerical scale
Mathematical operations are meaningful
Height, age, temperature, yield…
Nominal
No inherent order or ranking
I.E.: gender, race, color
Ordinal
Order series or ranking
I.E.: ranking table, scores,
Binary
Only two values
I.E.: Pass/Fail, Yes/No
Discrete
Countable, finite, only certain values are possible
I.E.: number of defects in particular lot or time frame…
Continuous
Infinitive possible values not countable
I.E.: temperature, pressure, humidity, length, time…
Limited by the precision of the measurement system (device)
Interval
Ordered, meaningful, equal differences between units
No natural zero
I.E.: temperature in F/C, IQ
Ratio
Ordered, menaingful, equal differences between units
Natural zero
I.E.: temperature in K, sales, weight, length...
Maximizing information

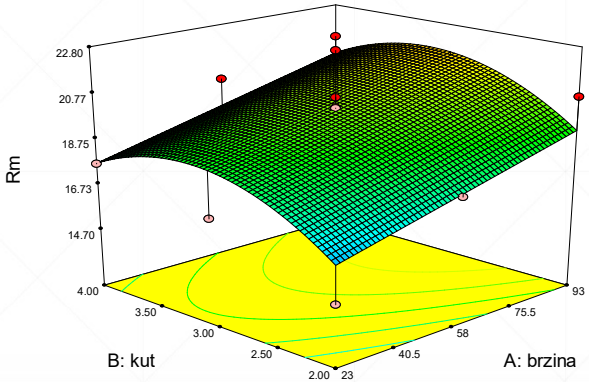
Results – effects, equation, response surface

Estimated Effects and Coefficients for Tvrdoca (coded units)

Term	Effect	Coef	SE Coef	T	P
Constant	586,417	0,8047	728,72	0,000	
A	25,500	12,750	0,8047	15,84	0,000
B	-14,000	-7,000	0,8047	-8,70	0,000
C	-4,333	-2,167	0,8047	-2,69	0,016
A*B	8,667	4,333	0,8047	5,38	0,000
A*C	0,333	0,167	0,8047	0,21	0,839
B*C	6,500	3,250	0,8047	4,04	0,001
A*B*C	3,167	1,583	0,8047	1,97	0,067

S = 3,94229 PRESS = 559,5
R-Sq = 95,99% R-Sq(pred) = 90,98% R-Sq(adj) = 94,24%

$$R_m = 70{,}02364 + 0{,}040111 \cdot v + 11{,}03252 \cdot \alpha_A - 17{,}04735 \cdot d_A - 1{,}72718 \cdot \alpha_A^2 + 1{,}04115 \cdot d_A^2$$



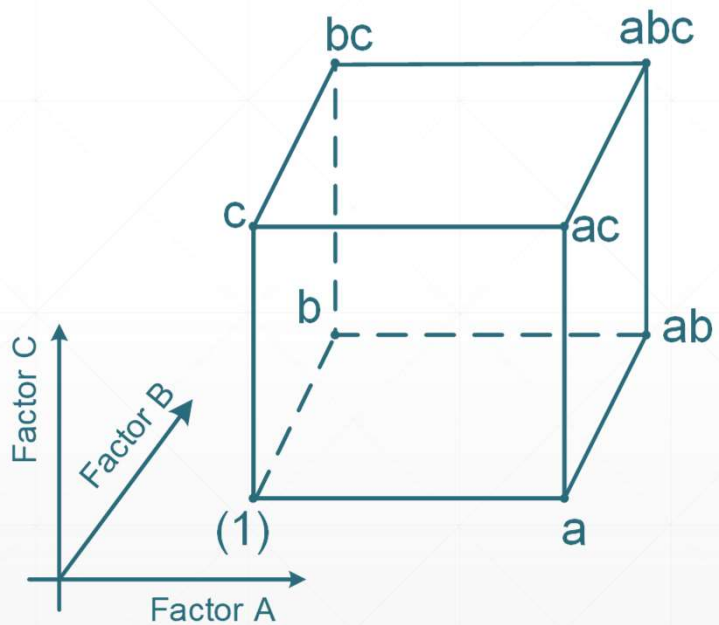
Analysis of effects

Final equation

Response surface

Factorial design

- Suitable for research the effect of 2, 3, 4, ... factors
- Selected factors can be examined at 2 or more levels simultaneously – provides the effect of interaction
- Provides pure effect of the given factor on the response
- Depending on the number of factor levels, the regression model will be composed of first and higher order members and interactions (first, second and higher order)



Evolution of factorial design $2^k r$

- OFAT (One Factor At Time) experiment:
  - Main effects – (impact on response by changing level of the factor)

    Effect A: $y_a - y_{(1)}$

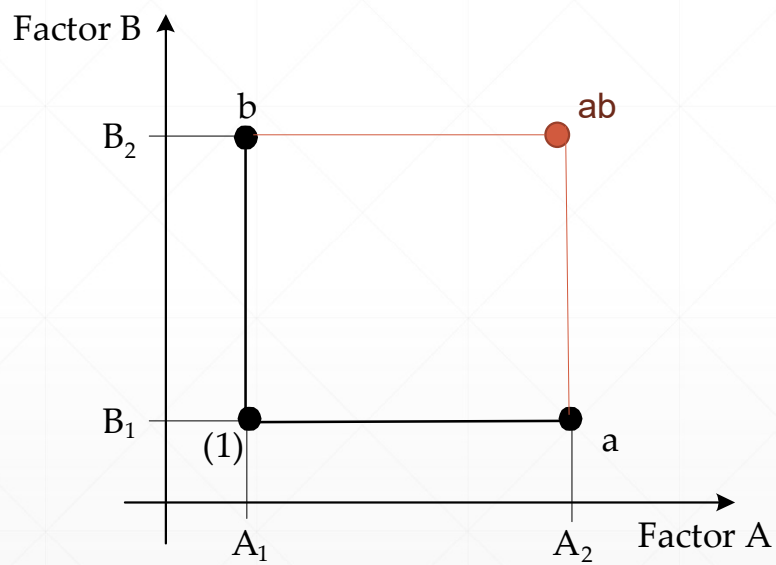
    Effect B: $y_B - y_{(1)}$

- Factorial design (2 level):
  - Main Effects - mean effect

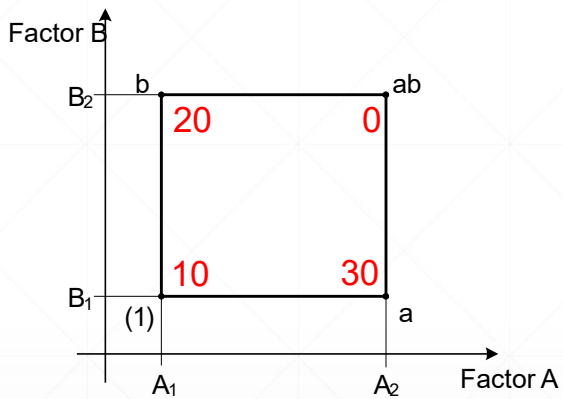
    Effect A: $\frac{(y_{ab}-y_b)+(y_a-y_{(1)})}{2}$

    Effect B: $\frac{(y_b-y_{(1)})+(y_{ab}-y_a)}{2}$

    Interaction AB: $\frac{(y_{ab}-y_b)-(y_a-y_{(1)})}{2}$ or $\frac{(y_b-y_{(1)})-(y_{ab}-y_a)}{2}$



Interaction and model complexity



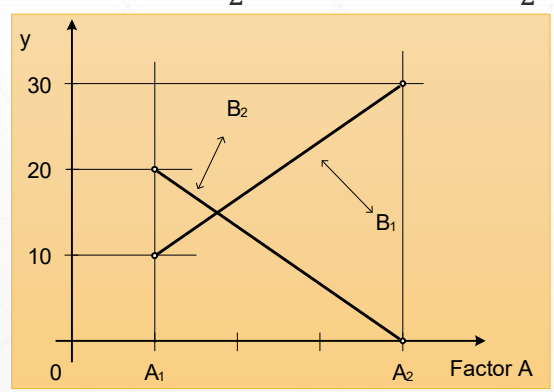
Main effects:

$$E_A = \frac{ab + a - b - (1)}{2} = \frac{0 + 30 - 20 - 10}{2} = 0$$

$$E_B = \frac{ab - a + b - (1)}{2} = \frac{0 - 30 + 20 - 10}{2} = -10$$

Interaction:

$$E_{AB} = \frac{(ab - a) - (b - (1))}{2} = \frac{0 - 30 - 20 + 10}{2} = -20$$



Plot of interactions

Regression model:

x_1- Factor A y – response variable
x_2- Factor B

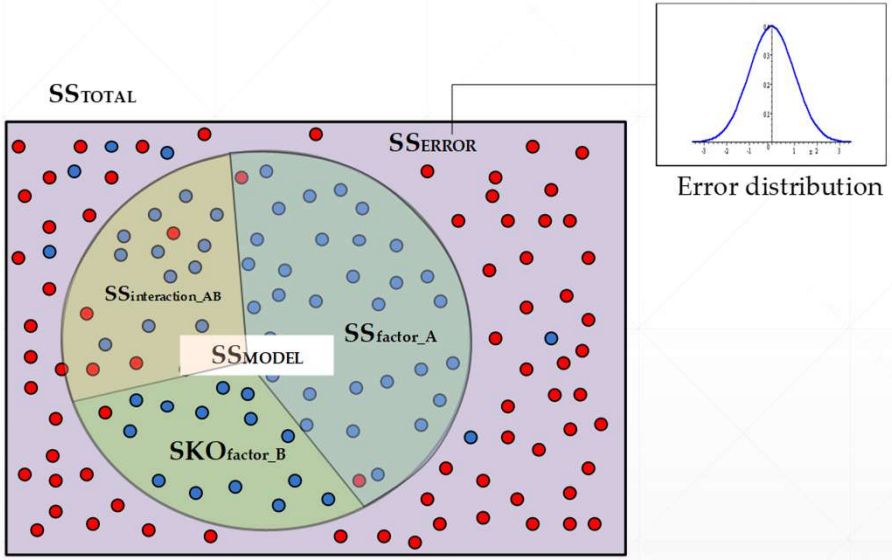
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

$$\hat{y} = 15 + 0 \cdot x_1 + \left(-\frac{10}{2}\right) x_2 + \left(\frac{-20}{2}\right) \cdot x_1 x_2$$

$$\hat{y} = 15 - 5x_2 - 10 \cdot x_1 x_2$$

Significance of the effects (terms)

- ANOVA - method for determination of statistical significance of the effects
- The errors should be normally distributed around 0 with inherent standard deviation
- Each variance is tested against the variance of the error (F-test)
- When p-value is smaller than 0.05 (5%) there is a statistical significance
  - p-value probability of false rejection of null hypothesis – probability that the difference in variation (F-value) is that large and is due to noise



ANOVA table

p<α – signif.
p>α – nonsignif.

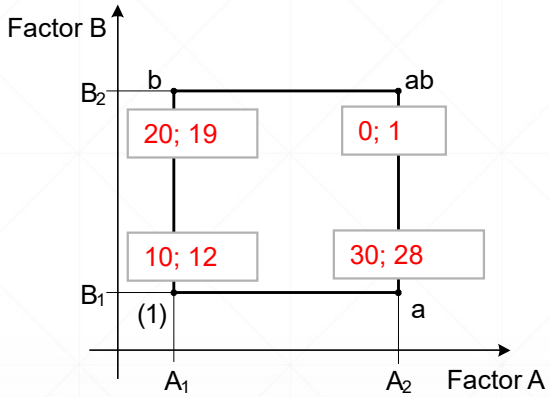
Main effects

2 factor interactions

3 factor interaction

Source	SS	*df*	MS	*F*	*p* -value
A	SS_A	$a - 1$	MS_A	MS_A/MS_{Error}	
B	SS_B	$b - 1$	MS_B	MS_B/MS_{Error}	
C	SS_C	$c - 1$	MS_C	MS_C/MS_{Error}	
AB	SS_{AB}	$(a - 1)(b - 1)$	MS_{AB}	MS_{AB}/MS_{Error}	
AC	SS_{AC}	$(a - 1)(c - 1)$	MS_{AC}	MS_{AC}/MS_{Error}	
BC	SS_{BC}	$(b - 1)(c - 1)$	MS_{BC}	MS_{BC}/MS_{Error}	
ABC	SS_{ABC}	$(a - 1)(b - 1)(c - 1)$	MS_{ABC}	MS_{ABC}/MS_{Error}	
$Error$	SS_{Error}	$abc(n - 1)$	MS_{Error}		

Example of ANOVA (2^2*2)



- Effects:

$$E_A = \frac{ab + a - b - (1)}{2} = -0.5$$

$$E_B = \frac{ab - a + b - (1)}{2} = -10$$

$$E_{AB} = \frac{(ab-a)-(b-(1))}{2} = -18.5$$

- Sum of squares:

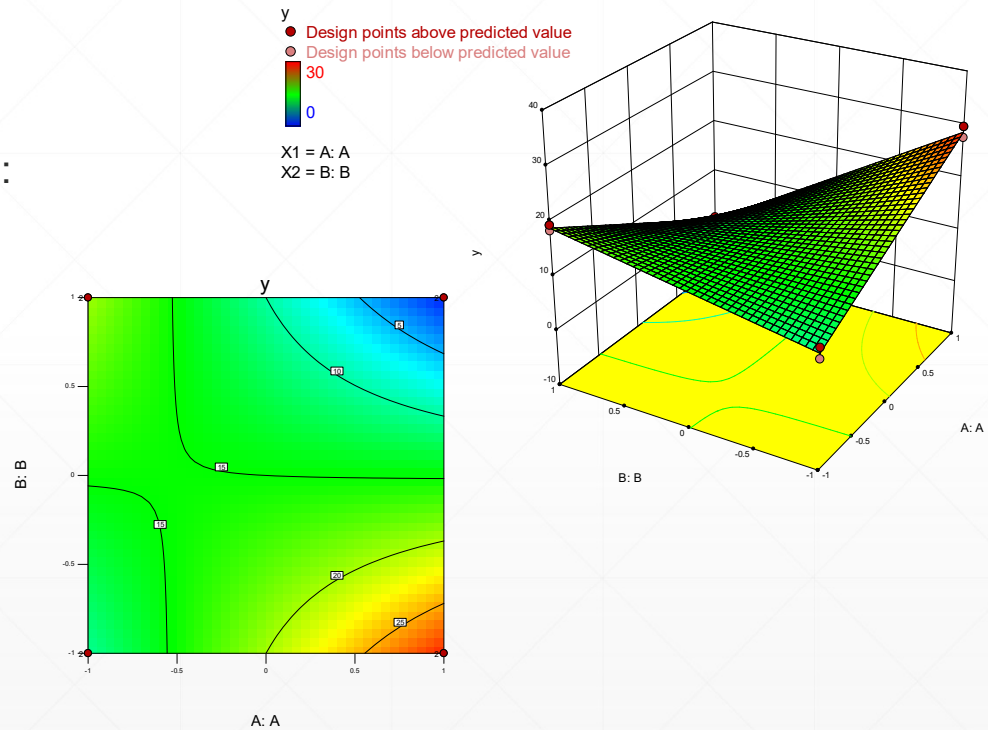
$$SS_A = r * E_A^2 = 2*0.25 = 0.5$$
$$SS_B = r * E_B^2 = 2*100 = 200$$
$$SS_{AB} = r * E_{AB}^2 = 2 * 342.25 = 684.5$$
$$SS_E = SS_T - SS_A - SS_B - SS_{AB} = 890$$

- ANOVA table

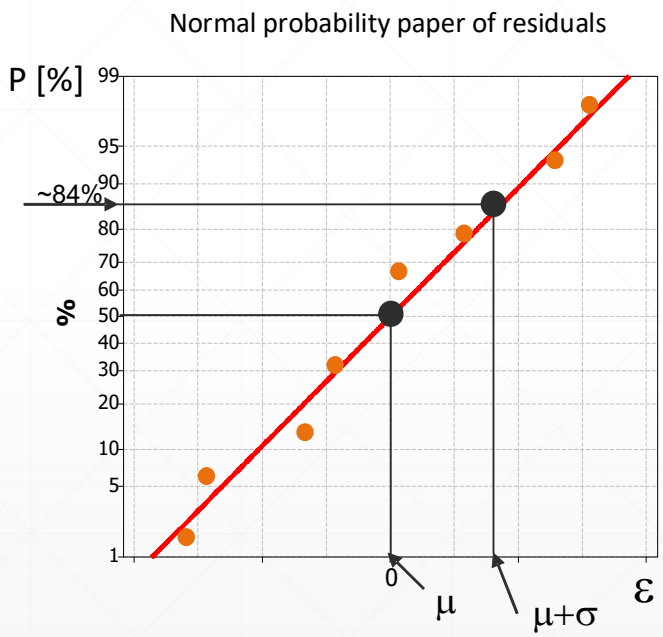
	Sum of		Mean	F	p-value
Source	**Squares**	**df**	**Square**	**Value**	**Prob > F**
Model	885.00	3	295.00	236.00	< 0.0001
A-A	*0.50*	*1*	*0.50*	*0.40*	*0.5614*
B-B	*200.00*	*1*	*200.00*	*160.00*	*0.0002*
AB	*684.50*	*1*	*684.50*	*547.60*	*< 0.0001*
Pure Error	5.00	4	1.25		
Cor Total	890.00	7			

Model

- General regression model for 2^2*2
  $$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$
- Regression model for the given example:
  $$\hat{y} = 15 - 0.25x_1 - 0.5x_2 - 9.25x_1x_2$$
- Contour plot and 3D response surface plot
- The model just mimic the phenomena which occurs in the process



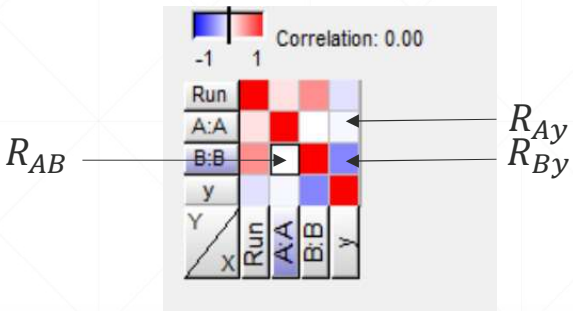
Properties of the model – residuals analysis



- Residuals – difference between actual and predicted values
- Analysis of residuals using normal probability paper
- Distribution of residuals:
  $$N\{\mu_\varepsilon = 0; \sigma_\varepsilon^2\}$$
- Residuals vs predicted analysis
- Residuals vs experiment run sequence
- Residuals vs factor

Properties of the model – R, R^2

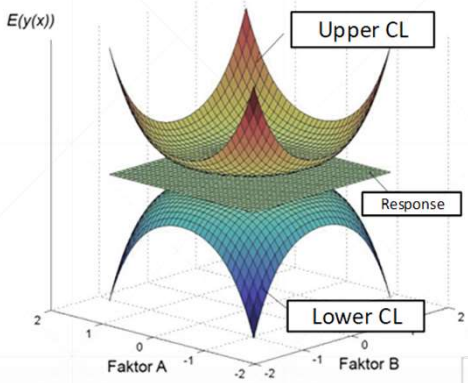
- Coefficient of correlation R – the factors (I.V.) must not be correlated
- Coefficient of determination R^2 – R-squared – the proportion of variance that is predictable from independent variables
- Adjusted R-squared – R^2 which is adjusted for the numbers of terms
- Predicted R-squared – the capability of model prediction when if data point is removed
- Parameters values should be close as possible to be able to state the adequacy of the model
- Need for more robust models

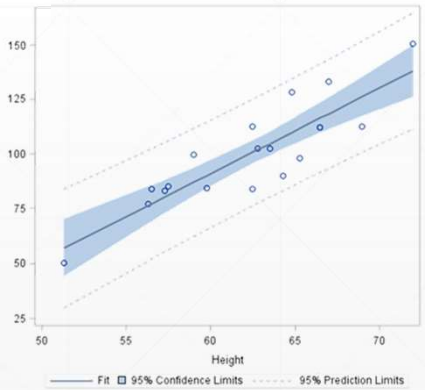


$$R^2 = 1 - \frac{SS_R}{SS_T}$$

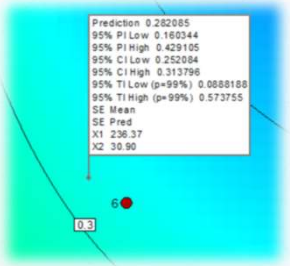
$$AdjR^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

The estimation and prediction error



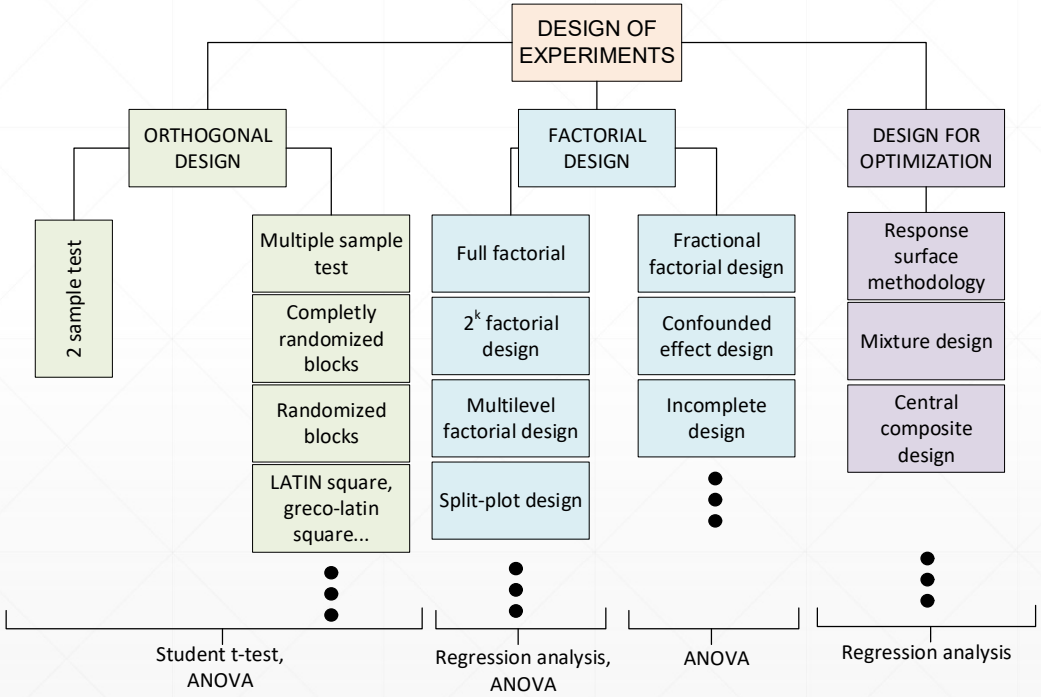


- Larger sample size better prediction (smaller prediction error) – sample theory (inference) and confidence interval construction
- CI or PI typically with 5% error (i.e. 95%CI or 95%PI)
- Estimation and prediction of the values must be provided by interval - not only one value!
- I.e. Confidence and prediction interval:

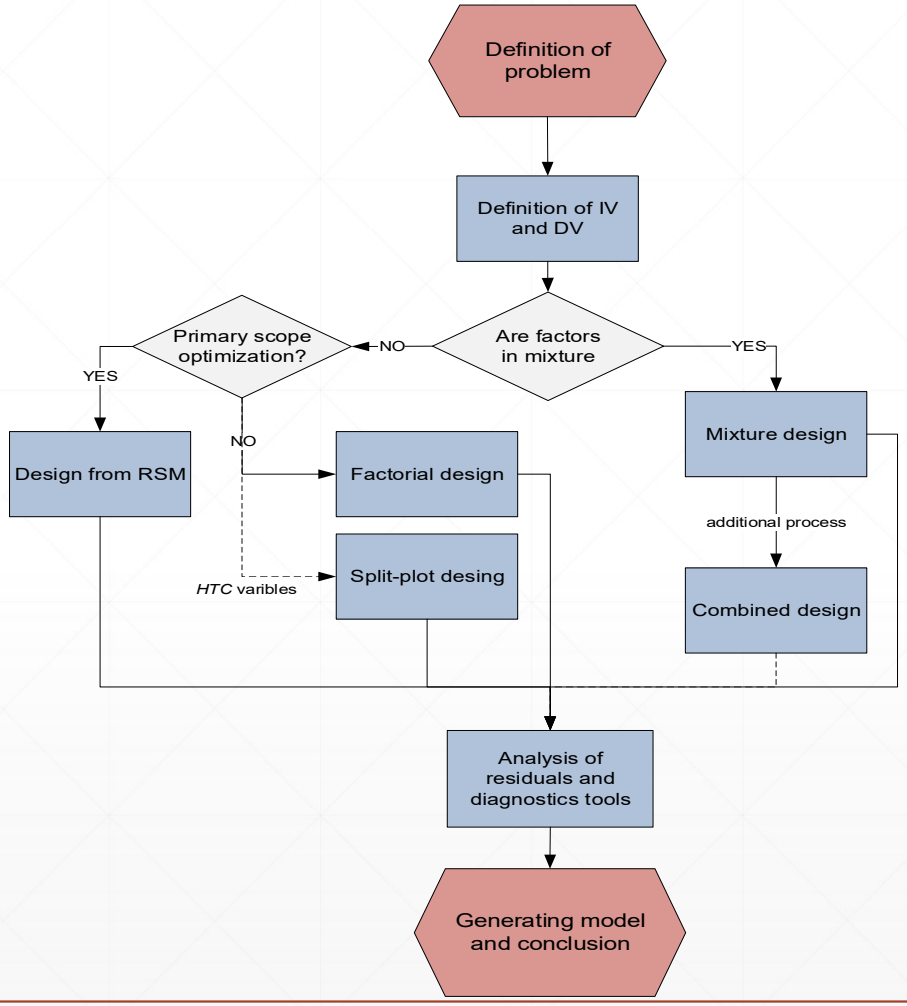


Common designs

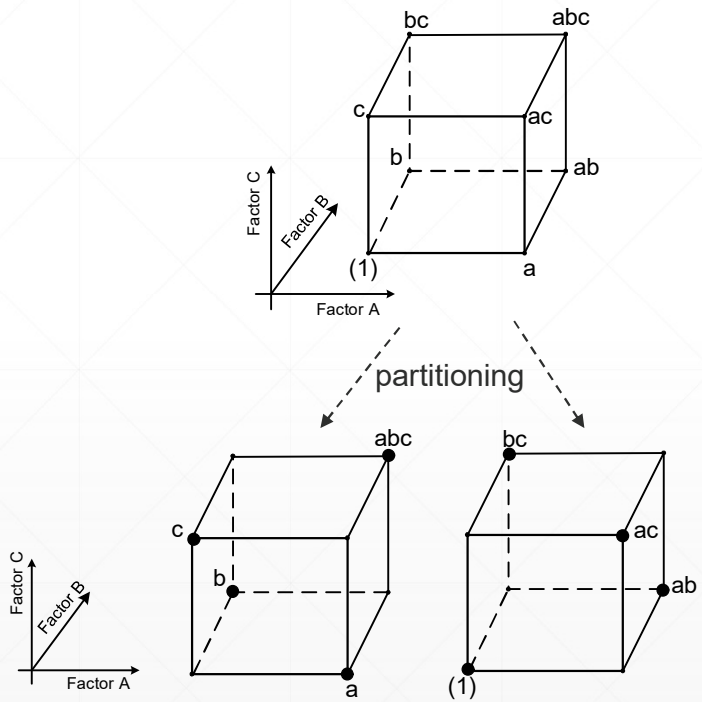
- The scope of the research directs the suitable design
  - Screening design
  - Effects estimation
  - Optimization
- Experimental area affect the choice of the design
- Cost-benefit analysis prior to design definition



DOE selection flowchart



Fractional factorial design



- To model cause-and-effect relationships between IV and DV more economical
- To determine the few important main effects from the many less important
- With assumption of non-significance of the high order interactions – Alias structure (i.e. 3 factor):

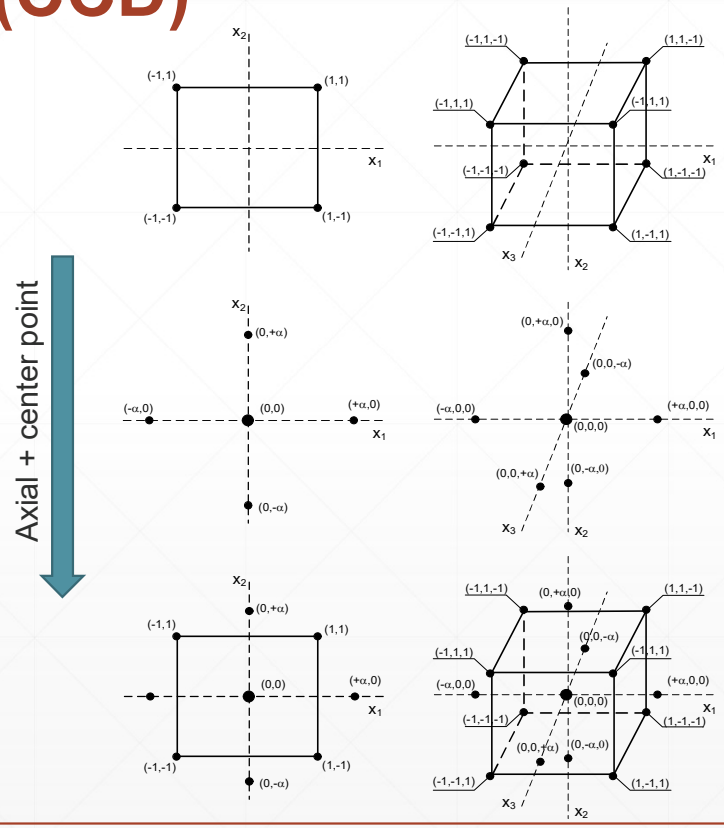
  A=A+BC
  B=B+AC
  C=C+AB
  I=I+ABC

- Resolution of the design 3, 4, 5 – related to the alias structure – higher resolution means that main effect are confounded with highest order interaction (most likely nonsignificant ones)

Central composite design (CCD)

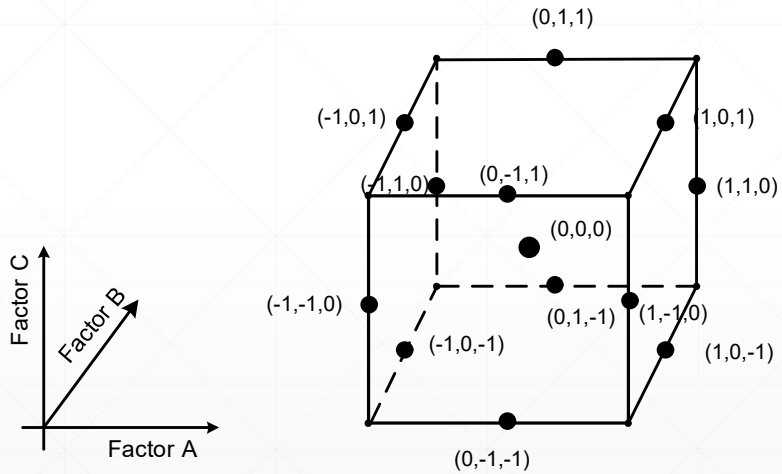
- The scope is to determine the regression equation (polynomial) – 2nd order model
- Basis of the CCD is 2 level factorial design
- Requires 5 l6evels of each factor
- α – distance of the axial points from center ($\alpha = \sqrt[4]{F}$- rotatable, $\alpha = 1$ - CCF design)
- It consists of factorial points, axial points and central point
- Less design points (experiments) than classical factorial design (i.e. 3^3 vs CCD)
- Regression model:

$$\hat{y} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum\sum_{j<i} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2$$



Box-Behnken design

- Quadratic design that doesn't include embedded factorial design
- 3 level of each factor with no extreme design points (such as in factorial design)
- BBD requires fewer treatment combinations than CCD - but poor prediction quality in some regions
- Better coefficient estimates near the center of the design space, but lack of prediction in corners of the cube (no design points there)
- Less robust than CCD due to sensitivity to missing points or mismeasured data



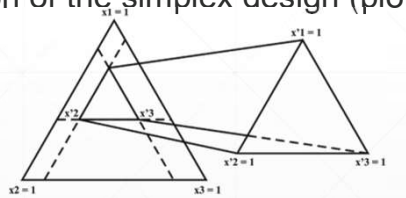
Mixture design

- Mixture consists of one or more components – sum of proportion is equal to 1

$$x_i > 0 \;\; i = 1, 2, 3 \;\; \dots q$$

$$\sum_{i=1}^{q} x_i = 1$$

- The construction of the simplex design (plot) – pseudo simplex design

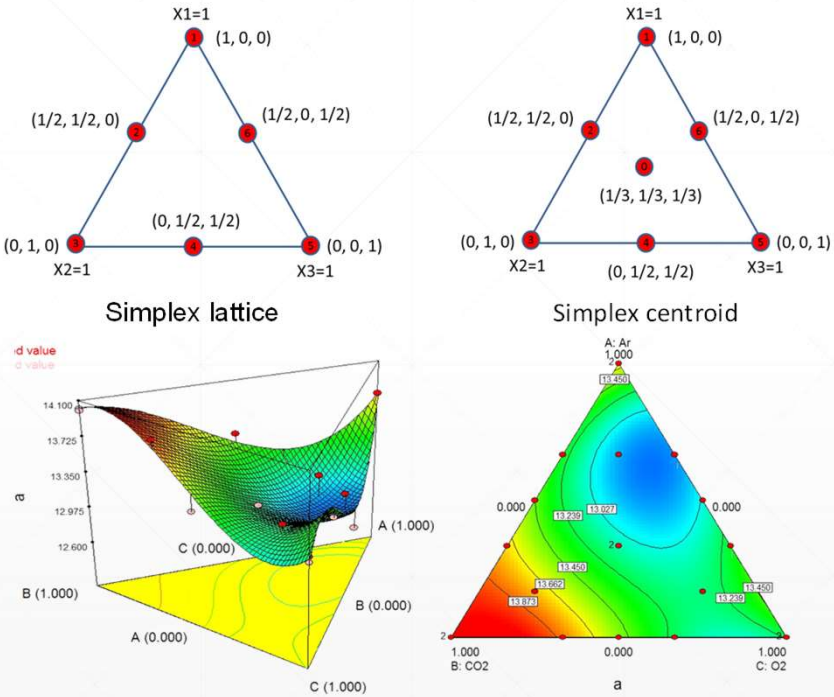


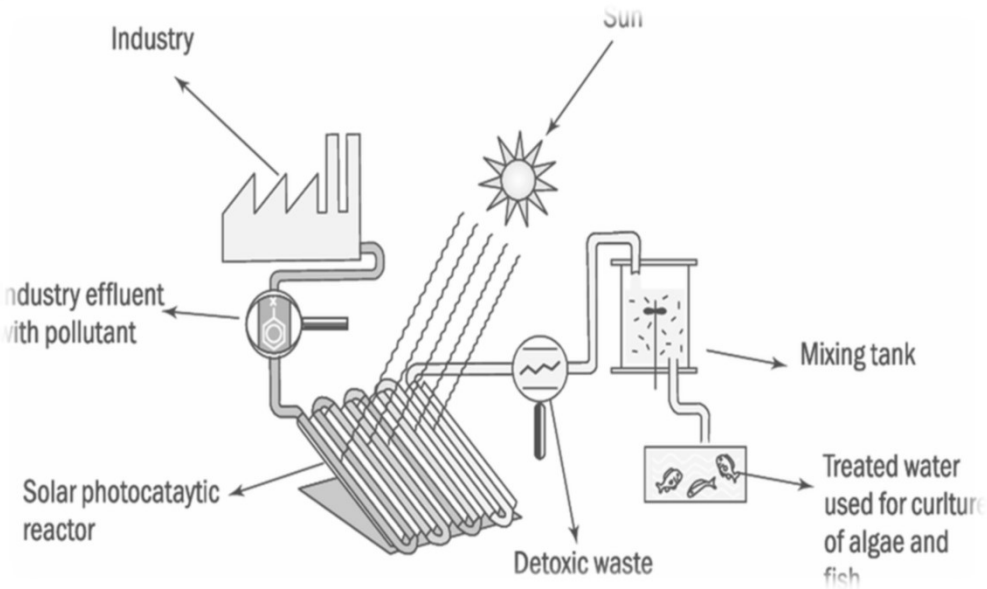
- Most common: simplex lattice, simplex centroid, simplex axial

- Regression model (quadratic):

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_3 + \beta_{12} x_1 x_2 + \beta_{23} x_2 x_3 + \beta_{13} x_1 x_3$$

$$(x_1^2 = x_1\,(1 - x_2 - x_3) = x_1 - x_1\,x_2 - x_1\,x_3)$$





SOURCE LINK: https://www.taylorfrancis.com/books/e/9780429058271/chapters/10.1201/9780429058271-20>

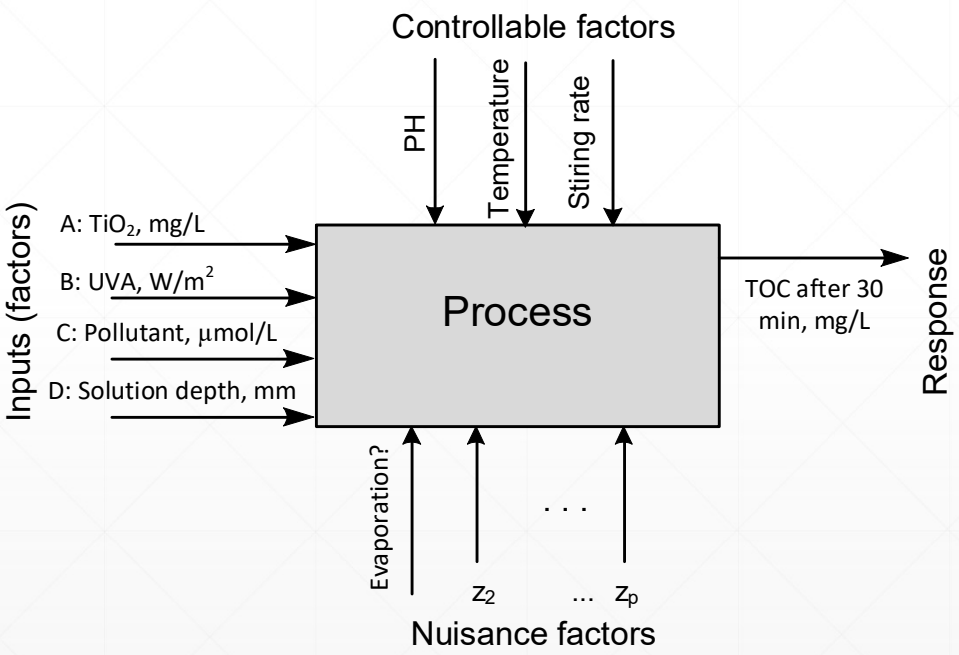
Case study – wastewater treatment by photocatalysis

Photolytic degradation of the color via UVA radiation in presence of TiO_2 with respect to conditions

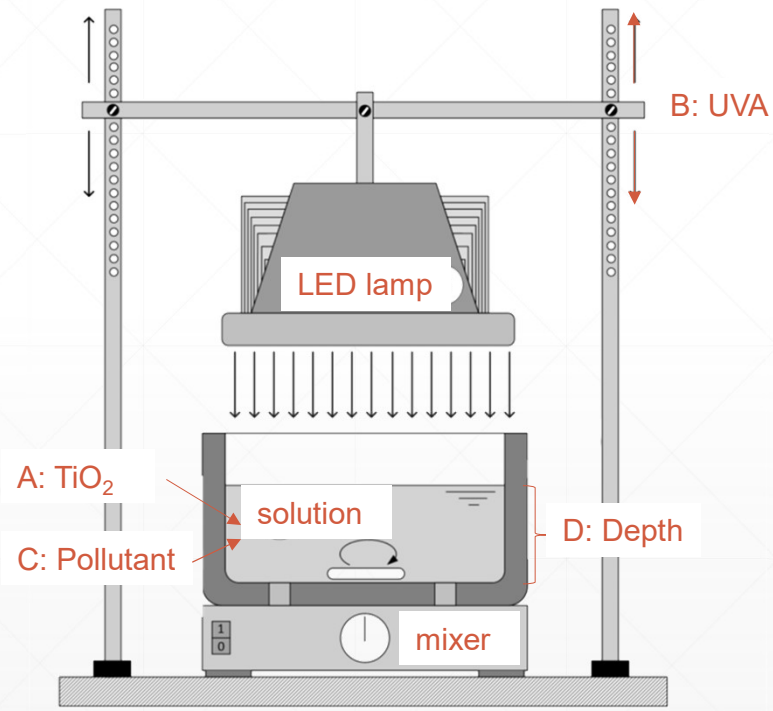
TOC (total organic carbon) in liquid samples – metrics of degradation

Definition of factors and responses

- The chosen input variables (factors):
  - Amount of TiO_2 in mg/L
  - Amount of UVA in W/m^2
  - Amount of pollutant in µmol/L
  - Water depth in mm
- Response variable – TOC after 30 min of reaction in mg/L
- Controllable factors:
  - PH of the samples
  - Sample temperature
  - Stiring rate



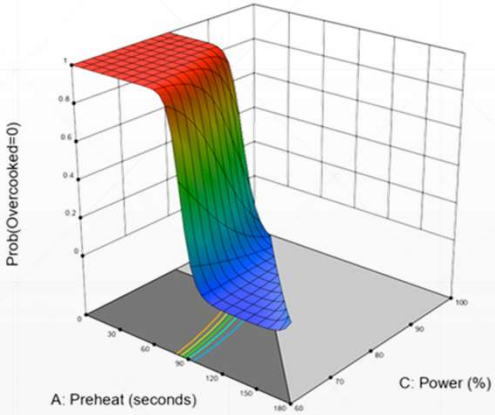
Experimental setup



- The most important fact - independently factor level changing
- Reproducibility of the factor setup
- Minimization of the noise induced by setup imperfections
- Assuring the possibility of random run of the design points
- Response variable – Amount of total organic carbon after 30 min in reaction – degradation rate
- Dye solution – simulate pollutant

Design expert software

- Designing experiments
- Analyzing data
- Visualizing results



Design definition and results

- Primary scope of the research - optimization of the degradation rate – RSM
- CCD design with 4 factors – 24 noncentral and 6 central points
- To assure rotatability feature α is set to 2
- Randomization – run order (run column) – minimizing the influence of the nuisance factors
- Factor levels:
  - A: (50, 450) mg/L
  - B: (10, 50) W/m^2
  - C: (10, 50) umol/L
  - D: (18.5, 32.5) mm

Std	Run	Factor 1 A:TiO2 mg/L	Factor 2 B:UVA W/m2	Factor 3 C:Polutant umol/L	Factor 4 D:Dubina mm	Response After 30min
1	16	450.00	30.00	30.00	25.50	0.12107
2	5	250.00	30.00	30.00	25.50	0.28597
3	19	350.00	40.00	40.00	29.00	0.32611
4	10	50.00	30.00	30.00	25.50	0.8167
5	30	150.00	20.00	40.00	22.00	0.86937
6	17	350.00	20.00	40.00	29.00	0.54171
7	9	150.00	40.00	20.00	29.00	0.21809
8	24	250.00	10.00	30.00	25.50	0.49753
9	28	150.00	20.00	40.00	29.00	0.83431
10	23	350.00	40.00	20.00	29.00	0.023171
11	6	150.00	40.00	40.00	29.00	0.78154
12	15	250.00	30.00	30.00	25.50	0.21514
13	*2*	*250.00*	*50.00*	*30.00*	*25.50*	*0.15073*
14	29	350.00	40.00	40.00	22.00	0.27831
15	13	150.00	20.00	20.00	22.00	0.14834
16	7	150.00	40.00	40.00	22.00	0.66984
17	14	250.00	30.00	50.00	25.50	1.008
18	12	250.00	30.00	30.00	25.50	0.2504
19	11	250.00	30.00	30.00	18.50	0.17739
20	20	250.00	30.00	30.00	25.50	0.21184
21	3	250.00	30.00	30.00	32.50	0.30535
22	21	250.00	30.00	30.00	25.50	0.23353
23	4	350.00	20.00	20.00	22.00	0.26253
24	27	250.00	30.00	30.00	25.50	0.39253
25	1	150.00	40.00	20.00	22.00	0.047904
26	22	150.00	20.00	20.00	29.00	0.17273
27	*26*	*350.00*	*20.00*	*40.00*	*22.00*	*1.0341*
28	18	250.00	30.00	10.00	25.50	0.003994
29	25	350.00	20.00	20.00	29.00	0.061878
30	8	350.00	40.00	20.00	22.00	0.0034628

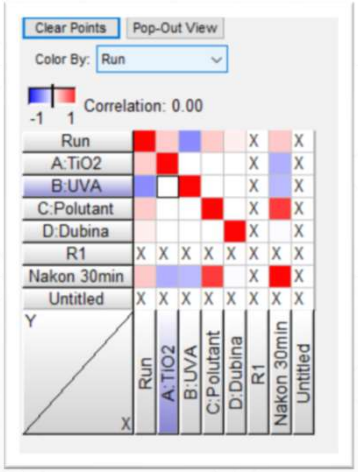
Design summary and evaluation

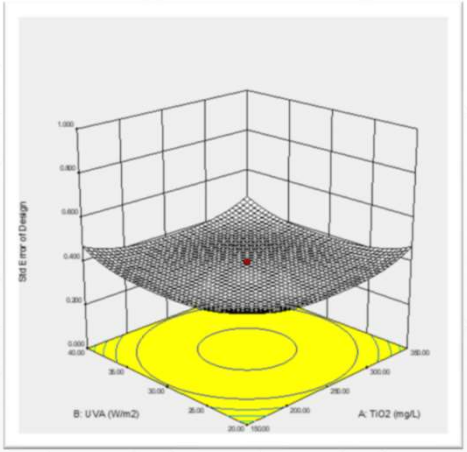
esign Summary

ile Versio	7.1.5.1		
tudy Type	Response Surface	Subtype	Randomized
esign Typ	Central Composite	Runs	30
esign Mo	Quadratic	Blocks	No Blocks

actor	Name	Units	Type	Subtype	Minimum	Maximum	Coded	Values	Mean	Std. Dev.
	TiO2	mg/L	Numeric	Continuous	50.00	450.00	-1.000=150.	1.000=350.0	250.00	90.97
	UVA	W/m2	Numeric	Continuous	10.00	50.00	-1.000=20.0	1.000=40.0	30.00	9.10
	Polutant	umol/L	Numeric	Continuous	10.00	50.00	-1.000=20.0	1.000=40.0	30.00	9.10
	Dubina	mm	Numeric	Continuous	18.50	32.50	-1.000=22.0	1.000=29.0	25.50	3.18

esponse	Name	Units	Obs	Analysis	Minimum	Maximum	Mean	Std. Dev.	Ratio	Trans
1	R1		0	Polynomial	No Data	No Data	No Data	No Data	N/A	None
2	Nakon 30min		30	Polynomial	0.0034628	1.0341	0.364786	0.309498	298.631	Square R
3	Untitled		0	Polynomial	No Data	No Data	No Data	No Data	N/A	None





Analysis of variance

- Due to non-homogeneity of the variance data transformation is needed

  Square root transformation: $y' = \sqrt{y + k}$

- Significant terms:
  - A, B, C, AB, AD, BD, A^2
- Nonsignificant terms:
  - D, AC, BC, CD, B^2, C^2, D^2, …
- Lack of fit – not significant

M	Intercept
M	A-TiO2
M	B-UVA
M	C-Polutant
	D-Dubina
M	AB
	AC
M	AD
	BC
M	BD
	CD
M	A^2
	B^2
	C^2
	D^2
	ABC
	ABD
	ACD
	BCD
	A^2B
	A^2C
	A^2D
	AB^2

Response 2 Nakon 30min

Transform: Square Roc Constant: 0

ANOVA for Response Surface Reduced Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	1.99	7	0.28	79.77	< 0.0001	significant
A-TiO2	0.22	1	0.22	62.97	< 0.0001	
B-UVA	0.18	1	0.18	51.31	< 0.0001	
C-Polutant	1.41	1	1.41	395.68	< 0.0001	
AB	0.056	1	0.056	15.67	0.0007	
AD	0.034	1	0.034	9.40	0.0057	
BD	0.060	1	0.060	16.94	0.0005	
A^2	0.023	1	0.023	6.45	0.0187	
Residual	0.079	22	3.570E-003			
Lack of Fit	0.059	17	3.468E-003	0.89	0.6178	not significa
Pure Error	0.020	5	3.918E-003			
Cor Total	2.07	29				

Effects analysis and model properties

Factor	Coefficient Estimate	df	Standard Error	95% CI Low	95% CI High	VIF
Intercept	0.52	1	0.014	0.49	0.55	
A-TiO2	-0.097	1	0.012	-0.12	-0.071	1.00
B-UVA	-0.087	1	0.012	-0.11	-0.062	1.00
C-Polutant	0.24	1	0.012	0.22	0.27	1.00
AB	-0.059	1	0.015	-0.090	-0.028	1.00
AD	-0.046	1	0.015	-0.077	-0.015	1.00
BD	0.061	1	0.015	0.031	0.092	1.00
A^2	0.028	1	0.011	5.181E-003	0.051	1.00

Std. Dev.	0.060	R-Squared	0.9621
Mean	0.54	Adj R-Squar	0.9500
C.V. %	10.99	Pred R-Squa	0.9115
PRESS	0.18	Adeq Precis	31.452
-2 Log Likelihood	-93.22	BIC	-66.01
		AICc	-70.37

- R-squared of 0.961 which is close to AdjR-sq of 0.95
- Factors A and B have negative effect thus C has positive:
  - When A (and B) has larger value the value of TOC is smaller
  - Factor C – higher amount of pollutant – higher TOC
- Factor D – the effect is nonsignificant – depth doesn't have any significant impact on TOC (considering domain)

Model – regression equation

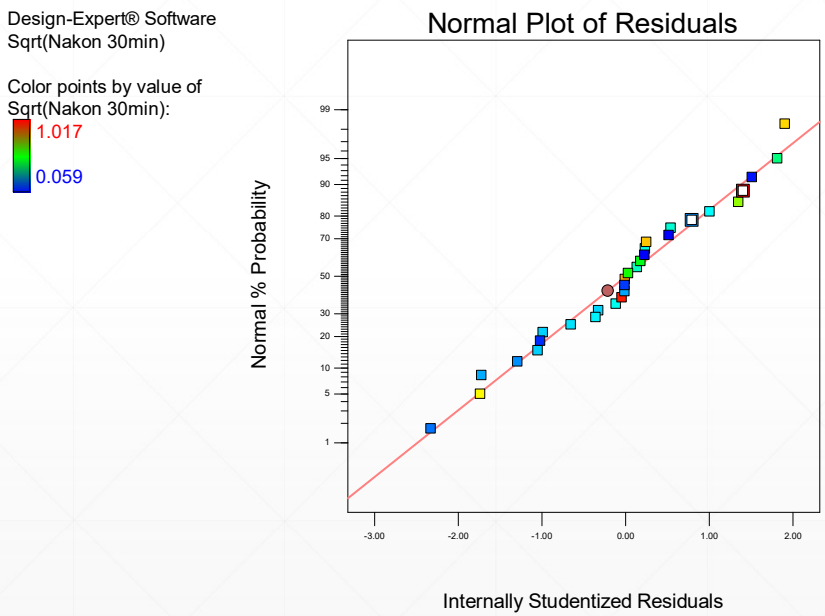
- The equation in coded factors:

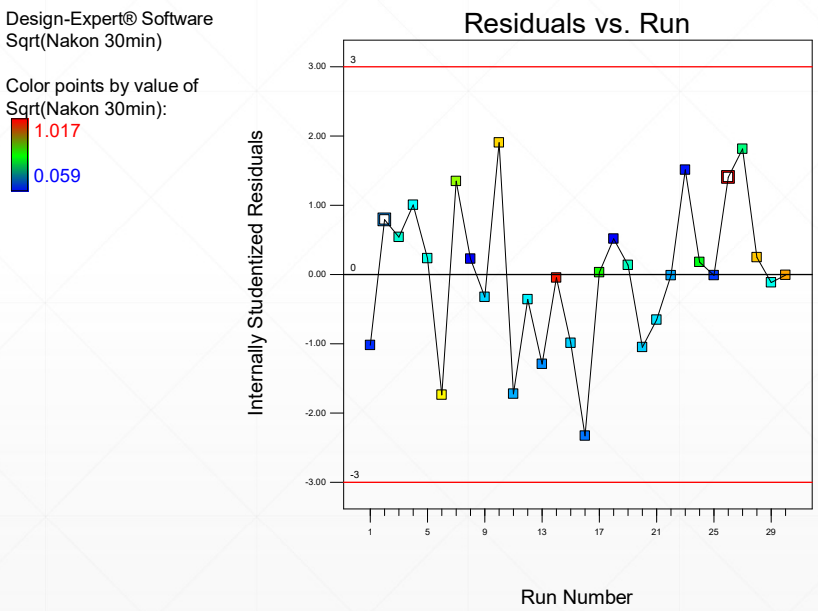
Sqrt (TOC 30min) = +0.52 - 0.097* A - 0.087* B + 0.24*C - 0.059*AB - 0.046*AD + 0.061* BD + 0.028* A^2

- The equation in actual values:

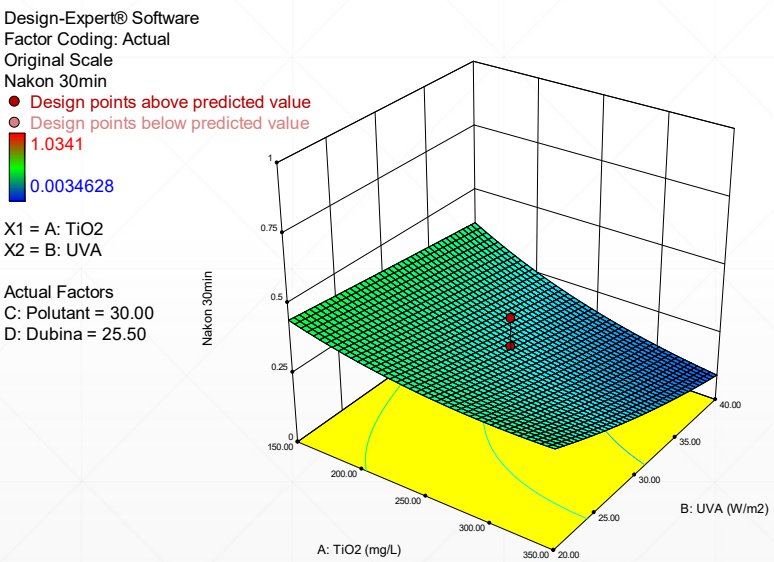
Sqrt(TOC 30min) =+0.030588 + 3.44405E-003*TiO2 - 0.030174*UVA + 0.024261*Pollutant - 5.91351E-005*TiO2*UVA - 1.58876E-004*TiO2*Depth +1.42044E-003*UVA*Depth + 2.82708E-006*$TiO2^2$

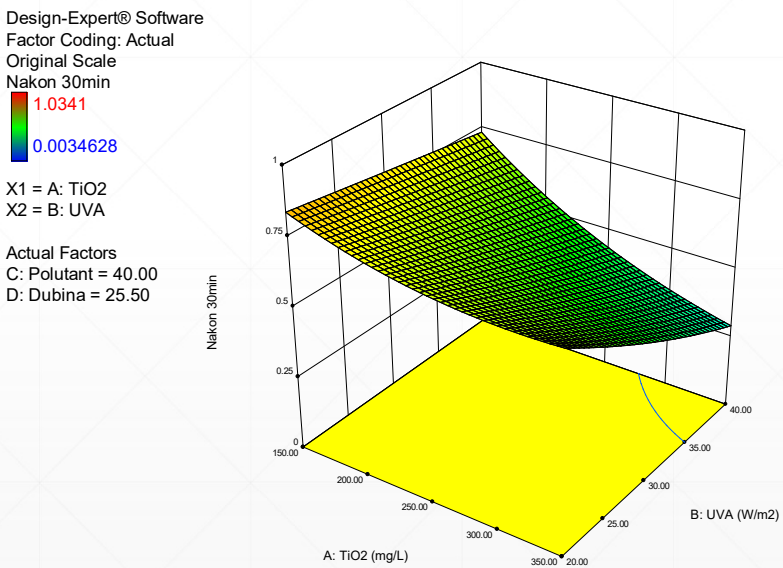
Diagnostics



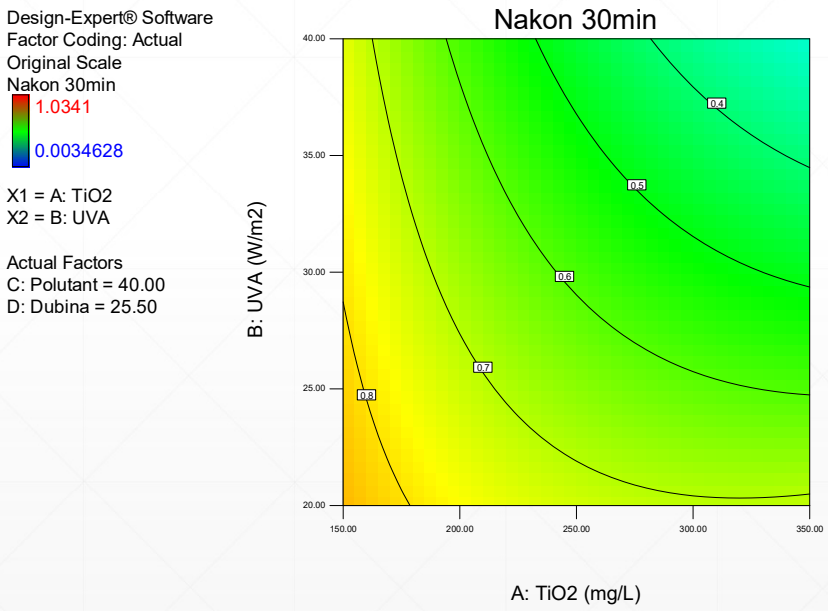


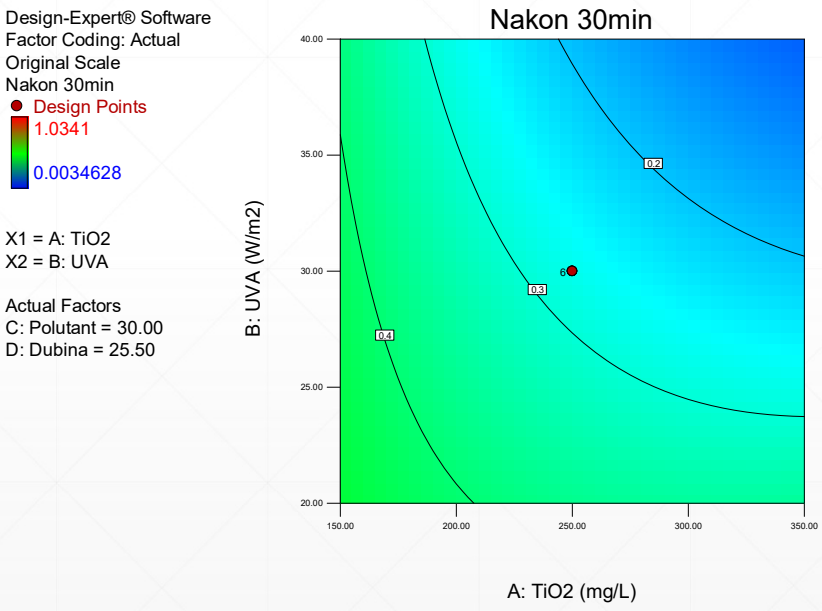
Graphically representation of model – 3D





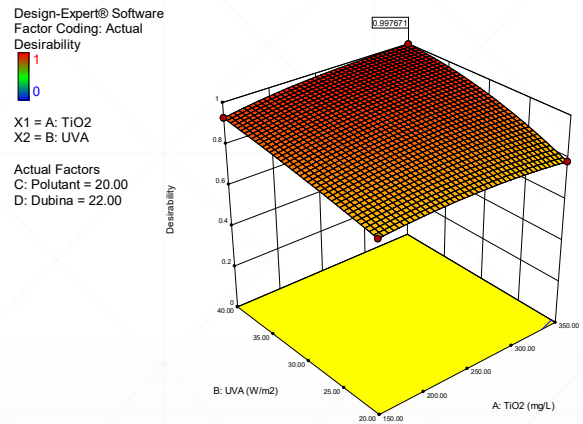
Graphically representation of model – contour diagram





Optimization of TOC

- Goal is to minimize pollutant residuals - TOC (after 30 min)
- Constraints (criteria) - boundaries which relates to technological and similar conditions
- Numerical or graphical optimization
- Criteria weights – relative importance especially when there is more than one response (goal)
- Desirability function – 0..1 (0-100%)



Constraints

Name	Goal	Lower Limit	Upper Limit	Lower Weight	Upper Weight	Importance
A:TiO2	is in range	150	350	1	1	1
B:UVA	is in range	20	40	1	1	3
C:Polutant	is in range	20	40	1	1	3
D:Dubina	is in range	22	29	1	1	3
Nakon 30min	minimize	0.0034628	1.0341	1	1	5

Solutions

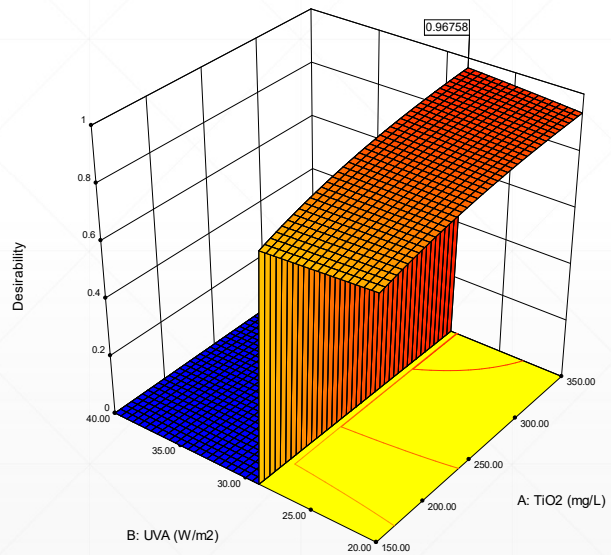
Number	TiO2	UVA	Polutant	Dubina	Nakon 30min	Desirability	
1	350.000	40.000	20.000	22.000	0.006	0.998	Selected
2	349.998	40.000	20.000	22.063	0.006	0.998	

Optimization of TOC („what if…”)

Design-Expert® Software
Factor Coding: Actual
Desirability
1
0

X1 = A: TiO2
X2 = B: UVA

Actual Factors
C: Polutant = 20.00
D: Dubina = 28.93



- Limitation: A<28 W/m2

Constraints

Name	Goal	Lower Limit	Upper Limit	Lower Weight	Upper Weight	Importance
A:TiO2	minimize	150	350	1	1	1
B:UVA	is in range	20	28	1	1	3
C:Polutant	is in range	20	40	1	1	3
D:Dubina	maximize	22	29	1	1	3
Nakon 30min	minimize	0.0034628	1.0341	1	1	5

Solutions

Number	TiO2	UVA	Polutant	Dubina	Nakon 30min	Desirability	
1	170.603	20.002	20.000	29.000	0.155	0.905	Selected
2	174.277	20.002	20.000	29.000	0.151	0.905	
3	175.301	20.000	20.000	29.000	0.150	0.905	
4	173.640	20.243	20.000	29.000	0.152	0.905	
5	176.406	20.266	20.000	29.000	0.149	0.904	

"Essentially, all models are wrong, but some are useful"

George Edward Pelham Box FRS
(1919 – 2013)