Some perspectives on Industrial Statistics with a view towards applications and research

by

John Tyssedal,
Trondheim
Symposium 2020

Outline

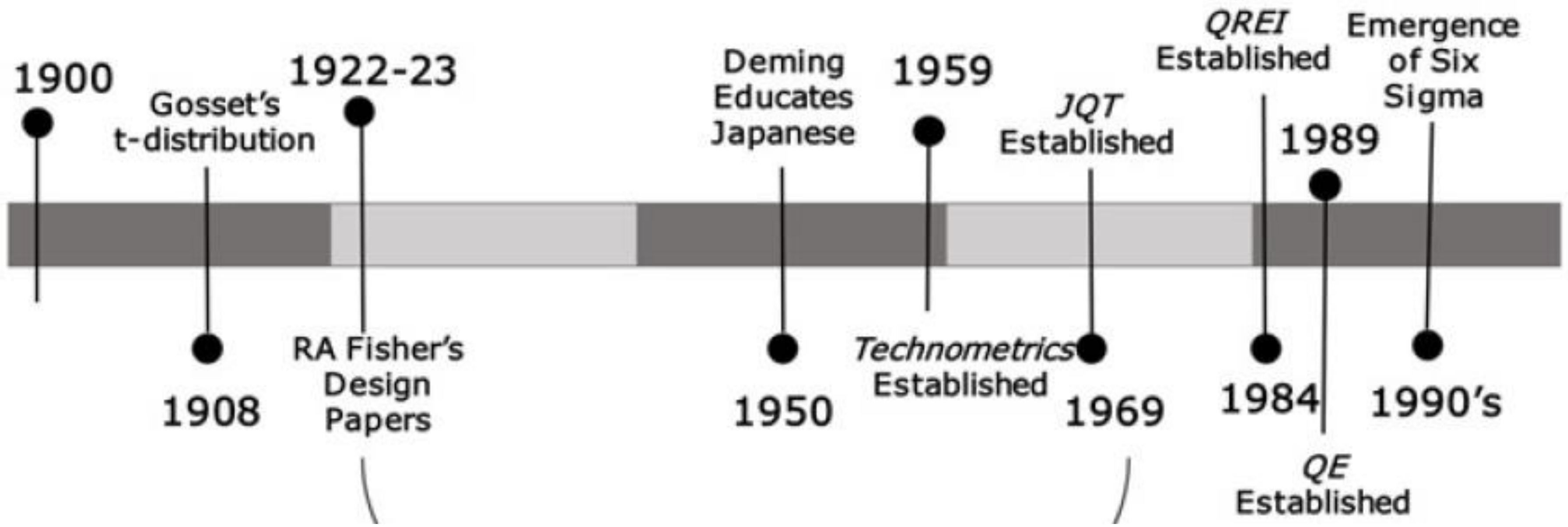
- Some history
- Some applications (highly subjective)
- Some research (highly subjective)
- Everything mixed.

WHAT IS INDUSTRIAL STATISTICS?

- Industrial Statistics is concerned with maintaining and improving the quality of goods and services. It involves a broad range of statistical tools but maintaining and improving quality involves an overall approach to the management of industrial processes that transcends the use of these specific tools. Variability is inherent in all processes, whether they be manufacturing processes or service processes. This variability must be controlled to create high quality goods and services and must be reduced to improve quality. Industrial Statistics focuses on the use of statistical thinking, i.e., the appreciation of the inherent variability of all processes. It also focuses on developing skills for modeling data and designing experiments that can lead to improvements in performance and reductions in variablity.

Department of Mathematics and Statistics, University of New Mexico

Timeline
Industrial Statistics



Statistical methods with roots in industry

- T-test - William Gosset, analyzing small samples of data at the Guinness Brewery
- The rank-sum test - Frank Wilcoxon, needed distribution free methods at American Cyanid
- Statistical process Control – Walter Shewart, monitoring and improving production at Bell Telephone Company
- Sequential probability ratio test – Abraham Wald, efficient munitions testing in World War II
- Ridge regression – Arhur Hoerl and Robert Kennard, problems with correlated predictors at Du Pont
- Exploratory data analysis – John Tukey, problems at Bell Laboratories
- Response surface methodology – George Box, problems at Imperial Chemical Industries


Pioneers in industrial statistics

- Jack Youden
- George Box
- Walter Shewart
- Stu Hunter

Industrial statistics in the 1970's

Ideal for an Industrial Statistician

Self-taught

Innovative Problem Solver

Clear Communicator

Open Sharer of knowledge

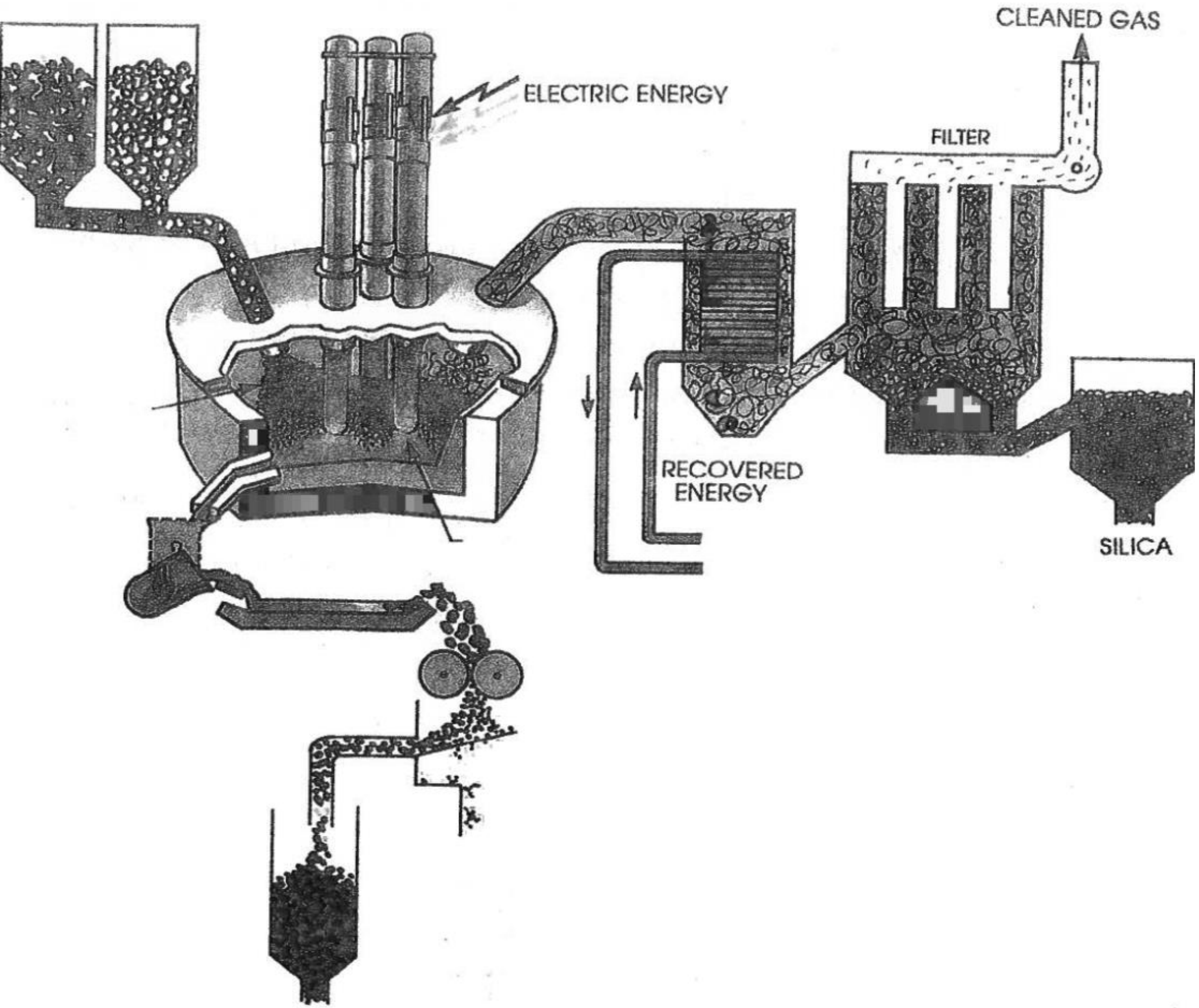
Emphasize on good study design

- Cambridge Dictionary defines «industry» as the companies and activities involved in the process of producing goods for sale.

- Dominating subjects
- DOE
- SPC
- Reliability

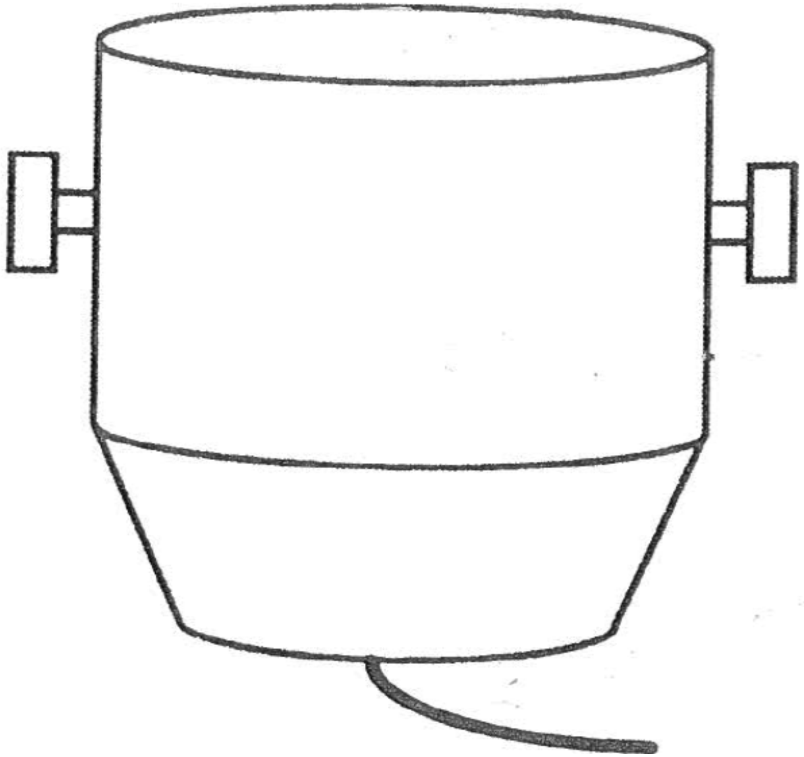
An Industrial Example

Silicon production at
Elkem Meråker



A scetch of Silicon Metal Production

Refining the silicon metal



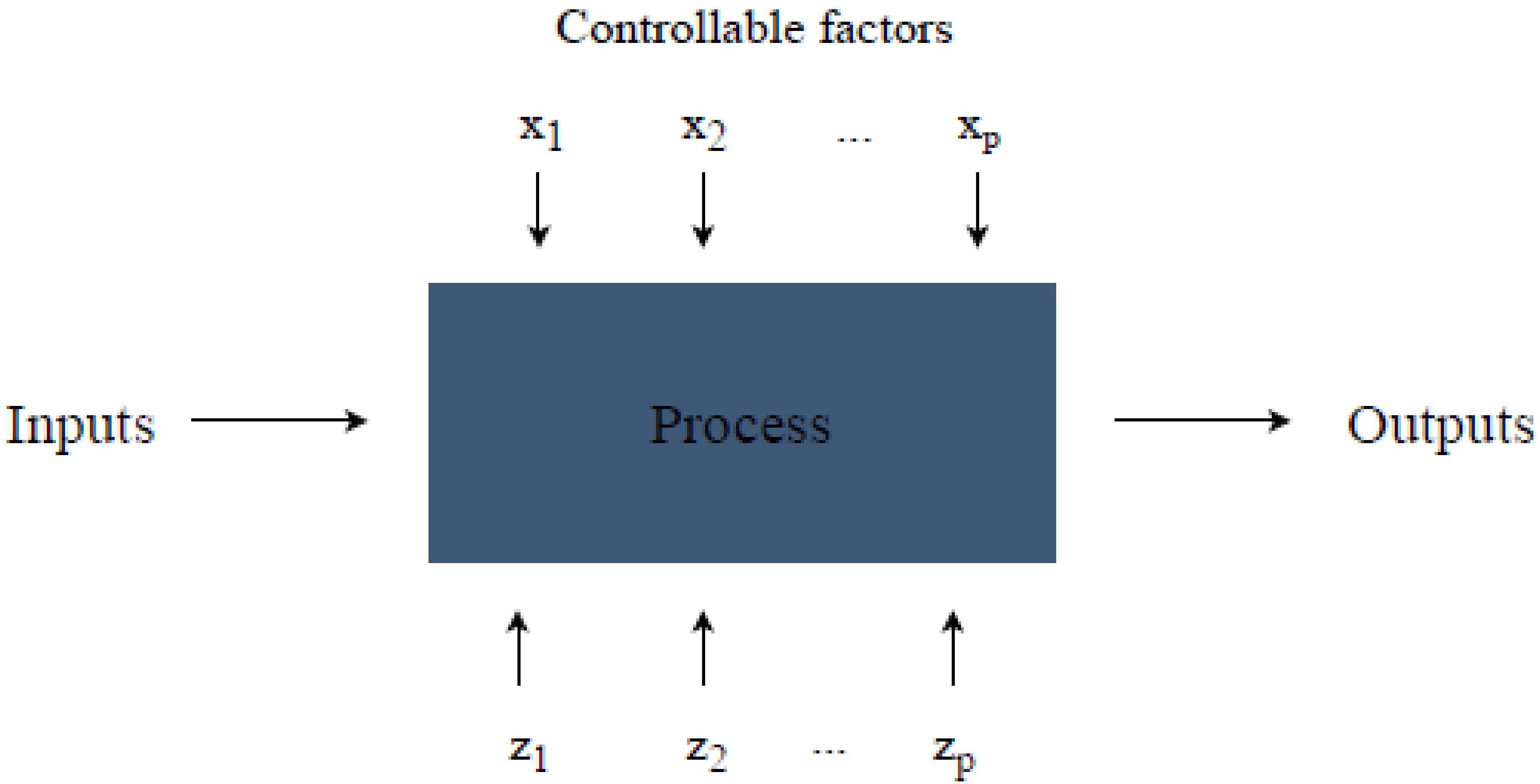
- Refining is done in the bailer used for tapping the furnace

Sand	Dolomitt	Gas	Refining time	Ca. change
-	-	-	-	y_1
-	+	-	+	y_2
+	-	+	-	y_3
+	+	+	+	y_4
+	-	-	+	y_5
+	+	-	-	y_6
-	-	+	+	y_7
-	+	+	-	y_8
0	0	0	0	y_9
0	0	0	0	y_{10}
0	0	0	0	y_{11}

- A 2^{4-1} experiment + 3 center runs in 3 blocks to reduce the calcium content in silicon.

Erlend Olsen 1996

A general model of a process



What happened to Industrial Statistics after 1970

Organizations with conferences

- ESRA
- ISBIS
- ENBIS

Other conferences

- Fall Technical
- QPRC
- ICISE
- MMR
- Spring Research

Movements

- Quality revolution
- Six-sigma

Dark clouds

- Manufacturing start moving from west to east.
- Easy to use software enabled engineers to handle routine statistical work themselves.
- Many statistical groups did not exist anymore

From the 2008 discussion in Technometrics about the future of statistics in industry:

- We live in a golden age of industrial statistics, but the status of statisticians in industry is possible at an all time low.

What happened to the main fields

SPC – From univariate to multivariate. From independent to correlated.

DOE – Screening experiments, optimal designs and computer experiments.

RELIABILITY – Life tests, degradation models, field data and Bayesian methods

Steinberg (2016)

Factor Screening

?

$$Y = f\left(x_1, x_2, \ldots, x_k\right) + \varepsilon$$

Factor screening is to identify those (normally few) factors out of potentially many that really affect the response (or explain most of the variation in the response).

It takes place at an early stage of experimentation when little or almost nothing is known.

Projectivity of two level designs

A $n \times k$ design with n runs and k factors each at two levels is said to be of projectivity P if the design contains a complete 2^p factorial in every possible subset of P out of k factors, possibly with some points replicated.

(Box and Tyssedal 1996)

An Industrial Example. Miocroplast Stjørdal 2003, now IV Moulding Leksvik

THE USE OF A 12 RUN PLACKETT-BURMAN DESIGN IN THE INJECTION MOULDING PRODUCTION OF A TECHNICAL PLASTIC COMPONENT

Microplast, Stjørdal 2003

- An injection moulding machine may have 15-20 variables that need to be set to operational conditions when production of a new product is started.
- Their strategy was one factor at a time experimentation and they realized the need for something that was more efficient.
- The leader of the project from SINTEF had taken a course in DOE

A 12 run PB design with 4 center runs

A	B	C	D	E	F	G	H	J	K	L	Y
1	-1	1	-1	-1	-1	1	1	1	-1	1	15,4
1	1	-1	1	-1	-1	-1	1	1	1	-1	17,3
-1	1	1	-1	1	-1	-1	-1	1	1	1	19,3
1	-1	1	1	-1	1	-1	-1	-1	1	1	17,4
1	1	-1	1	1	-1	1	-1	-1	-1	1	21,3
1	1	1	-1	1	1	-1	1	-1	-1	-1	19,3
-1	1	1	1	-1	1	1	-1	1	-1	-1	17,3
-1	-1	1	1	1	-1	1	1	-1	1	-1	21,4
-1	-1	-1	1	1	1	-1	1	1	-1	1	21,3
1	-1	-1	-1	1	1	1	-1	1	1	-1	19,4
-1	1	-1	-1	-1	1	1	1	-1	1	1	15,3
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	15,3
0	0	0	0	0	0	0	0	0	0	0	18,4
0	0	0	0	0	0	0	0	0	0	0	18,3
0	0	0	0	0	0	0	0	0	0	0	18,3
0	0	0	0	0	0	0	0	0	0	0	18,4

- Factors A-H represent: - Pressure, two velocity factors, two time factors, three temperature factors.
- Columns J-L represent unassigned factors.
- The response Y is here cycle time.
- The other responses measured represented weight and several length and width measures in order to meet specification limits. One of the responses was derived from two of the others.

The result of a Projective Based Search for Cycle Time

1 ACTIVE		2 ACTIVE		3 ACTIVE	
$\hat{\sigma}$	FAKTOR	$\hat{\sigma}$	FAKTOR	$\hat{\sigma}$	FAKTOR
1,10	E	0,06	D,E	0,05	B,D,E
2,19	D	1,10	E	0,05	D,E,J
2,34	-	1,15	A,E	0,05	D,E,L
2,45	B	1,15	C,E	0,06	C,D,E
2,45	A	1,15	E,G	0,06	A,D,E

- Conclusion: Factor D and E were responsible for most of the variation in the data.

Hallgeir Grinde

Non-Regular Designs

Advantages

Flexible run sizes

Good projection properties

Only partial aliased effects (for most of them)

How to analyze them?

Regularization techniques

- Lasso: $\min\limits_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^k} ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_{l_2}$ subject to $||\hat{\boldsymbol{\beta}}||_{l_1} \leq t$

Tibshirani (1996)

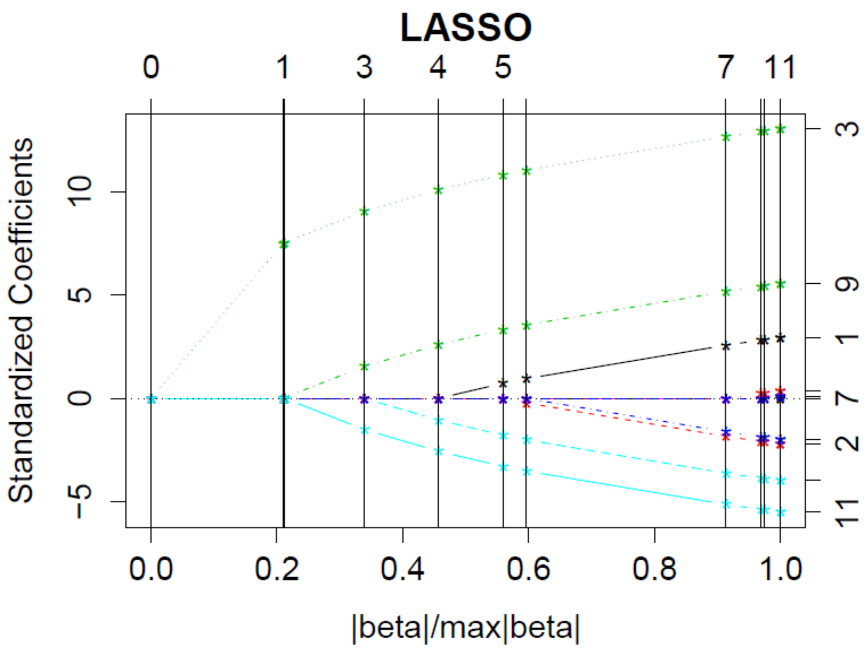
- Dantzig selector: $\min\limits_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^k} ||\hat{\boldsymbol{\beta}}||_{l_1}$ subject to $||\boldsymbol{X}^T \boldsymbol{r}||_{l_\infty} \leq \delta$

Candes and Tao (2007)

Analyzing a 12 run (PB12) non-regular designs

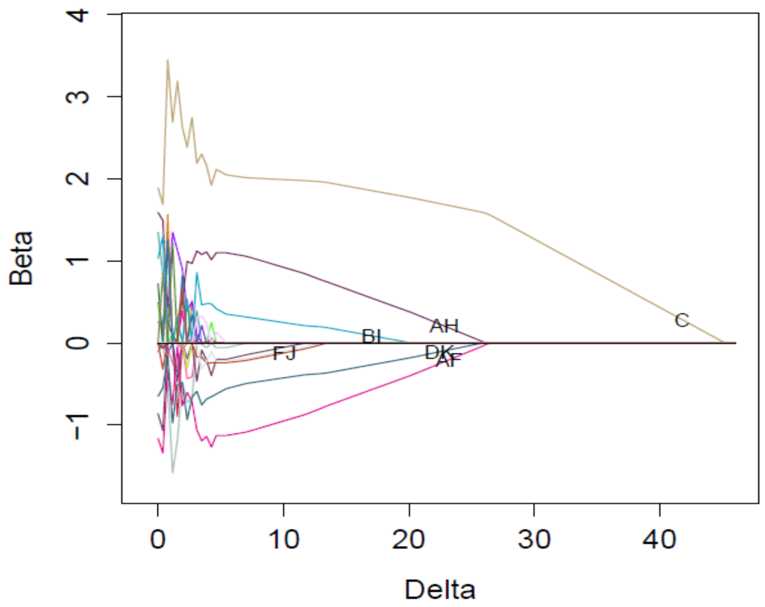
A	B	C	D	E	F	G	H	I	J	K	y_{1i}
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0.16
−1	−1	−1	−1	−1	1	1	1	1	1	1	0.49
−1	−1	1	1	1	−1	−1	−1	1	1	1	4.11
−1	1	1	1	1	−1	1	1	−1	−1	1	-8.32
−1	1	−1	−1	1	1	−1	1	−1	1	−1	4.01
−1	1	1	1	−1	1	1	−1	1	−1	−1	7.71
1	−1	1	1	−1	−1	1	1	−1	1	−1	8.36
1	−1	−1	−1	1	1	1	−1	−1	−1	1	4.07
1	−1	1	1	1	1	−1	1	1	−1	−1	-0.18
1	1	−1	−1	−1	−1	−1	1	1	−1	1	8.16
1	1	1	1	−1	1	−1	−1	−1	1	1	-4.24
1	1	−1	−1	1	−1	1	−1	1	1	−1	0.00

Variable selection with Lasso and the Dantzig selector on 11 main effects and 55 two-factor interactions



C, AF, AH,DK,BI,FJ

- **Dantzig selector**

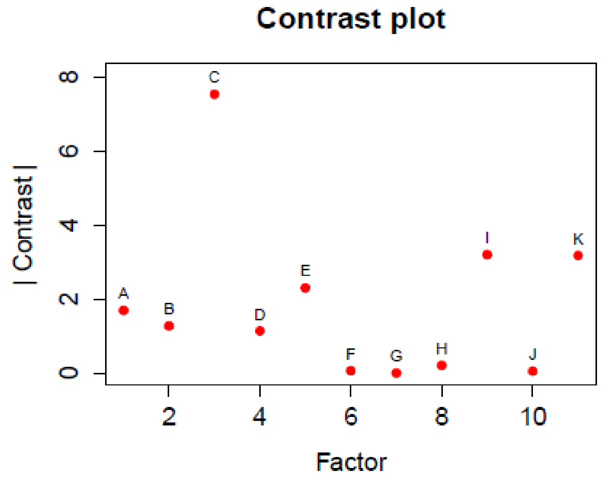


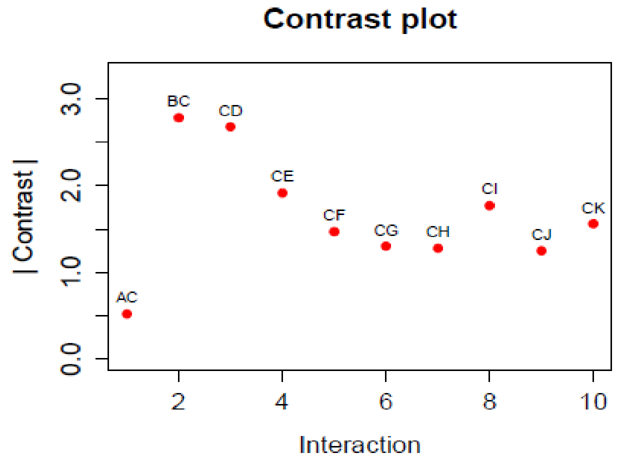
C,AF,AH,DK,BI,FJ

Wiik(2014)

Projection based analysis and contrast plot

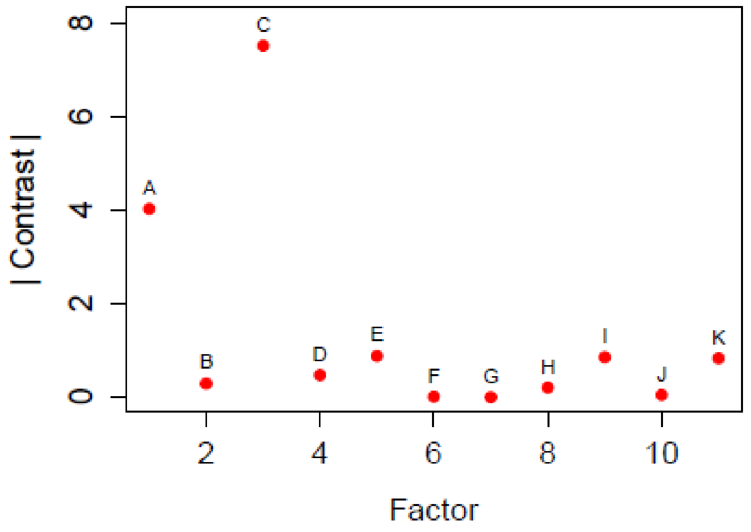
1 active		2 active		3 active	
Factor	$\hat{\sigma}^2$	Factors	$\hat{\sigma}^2$	Factors	$\hat{\sigma}^2$
C	9.44	C,I	6.79	A,C,E	0.04
E	23.39	C,K	7.10	C,E,I	1.83
I	23.43	B,C	8,29	C,E,K	2,16
K	24.87	C,E	8,43	C,F,H	3,08



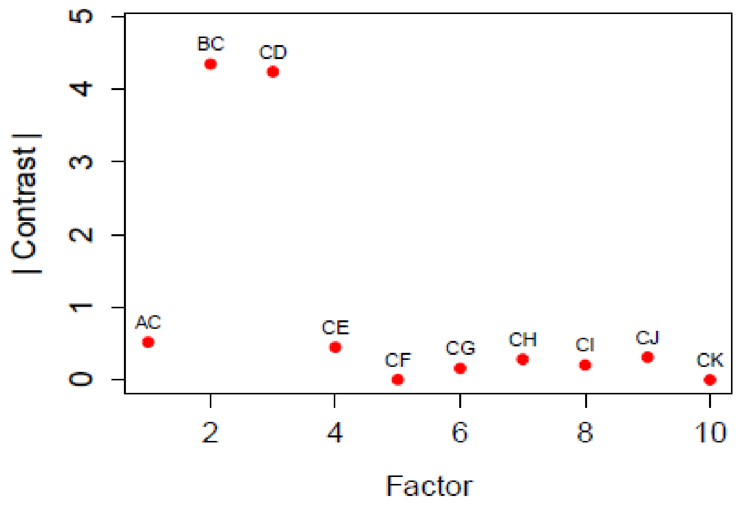


Alias reduced contrast plots

Main effects



- Two-factor interactions with C



Alias reduction

Suppose $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_1\boldsymbol{\beta}_1$

Then $E(\mathbf{Y}) = \mathbf{X}(\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\beta}_1) + (\mathbf{X}_1 - \mathbf{X}\mathbf{A})\boldsymbol{\beta}_1$

Let $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_1$, then $(\mathbf{X}_1 - \mathbf{X}\mathbf{A}) \perp \mathbf{X}$

Thereby $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\beta}_1$ can be estimated unbiased

Tyssedal and Niemi (2014)

Why can two-level experiments be so successful?

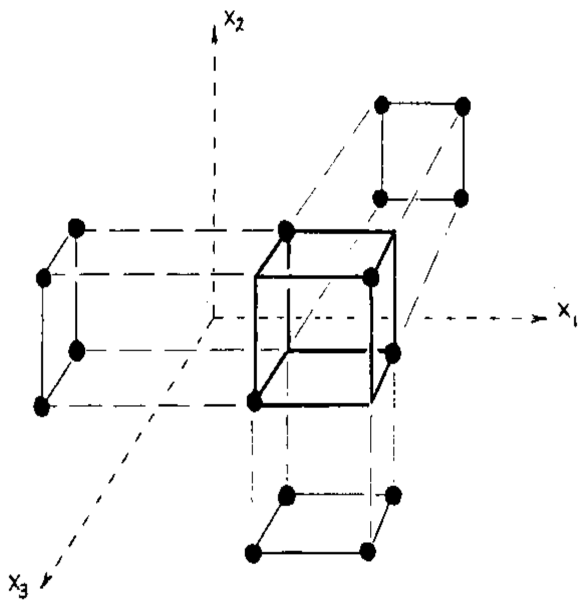


FIGURE 2—Projection of 2^{3-1}_{III} into three 2^2 factorials.

- George Box: I have always believed that the success of the two-level designs is due to their projection properties.
- BHH: Block what you can and randomize what you cannot.

Projectivity of blocked two level designs

A blocked two-level design is said to be of projectivity P or P_{α} if for any selection of P columns of the design all factorial effects up to and including P-factor interactions or α-factor interactions are estimable respectively.

(Hussain and Tyssedal 2016)

Blocking the 2_{IV}^{8-4} design in two blocks

- The 2_{IV}^{8-4} design

A	B	C	D	E	F	G	H
-1	-1	-1	-1	-1	-1	-1	-1
1	-1	-1	-1	1	1	1	-1
-1	1	-1	-1	1	1	-1	1
1	1	-1	-1	-1	-1	1	1
-1	-1	1	-1	1	-1	1	1
1	-1	1	-1	-1	1	-1	1
-1	1	1	-1	-1	1	1	-1
1	1	1	-1	1	-1	-1	-1
-1	-1	-1	1	-1	1	1	1
1	-1	-1	1	1	-1	-1	1
-1	1	-1	1	1	-1	1	-1
1	1	-1	1	-1	1	-1	-1
-1	-1	1	1	1	1	-1	-1
1	-1	1	1	-1	-1	1	-1
-1	1	1	1	-1	-1	-1	1
1	1	1	1	1	1	1	1

- E=ABC, F=ABD, G=ACD, H=BCD
- Every projections onto 3 factors is a replicated 2^3 design.

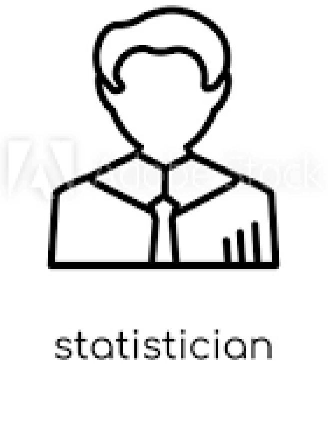
- Recommended block factor: AB.
- AB=CE=DF=GH
- 24 out 56 projections does not allow the estimation of all two-factor interactions.
- When blocked it is a *P*=1 design or a (16,8,1,2) screen.
- Only four factors are allowed in a 16 run regular two-level design in order to have a *P*=3 design when it is blocked the recommended way.

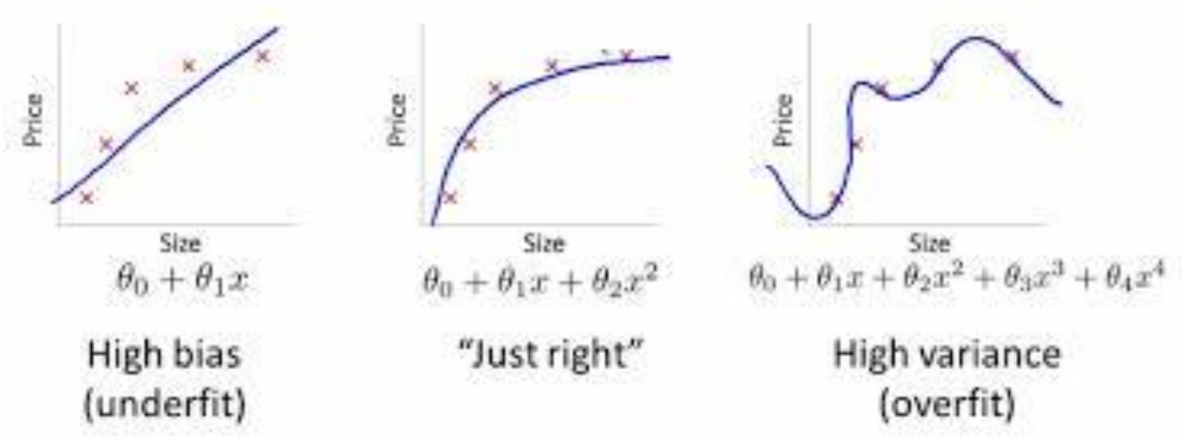
Some collected results for blocking

Design	P	R-Blocked	Strategy	Screen	Min D_s	Ave D_s	Max D_s
2_{IV}^{8-4}	3	(16,8,1,2)	MIP	(16,8,3,2)	0.917	0.929	1
2_{V}^{5-1}	4	(16,5,1,2)	Had+All	(16,5,3,2)	0.917	0.934	1
2_{IV}^{16-11}	3		MIP	(32,16,3,2)	0.917	0.970	1
2_{IV}^{16-11}	3		MIP	(32,16,3,4)	0.834	0.908	1
2_{VI}^{6-1}	5	(32,6,2,2)	MIP	(32,6,4,2)	0.917	0.959	0.982
2_{VI}^{6-1}	5	(32,6,1,4)	MIP	(32,6,4,4)	0.826	0.870	0.924
2_{IV}^{7-2}	3	(32,7,2,2)	Had	(32,7,3,2)	0.917	0.982	1
2_{IV}^{7-2}	3	(32,7,1,4)	Had	(32,7,3,4)	0.917	0.939	1
2_{V}^{8-2}	4	(64,8,2,4)	Had	(64,8,4,4)	0.917	0.966	1
2_{V}^{8-2}	4	(64,8,1,8)	Had	(64,8,4,8)	0.808	0.889	0.917

Hamre (2018),

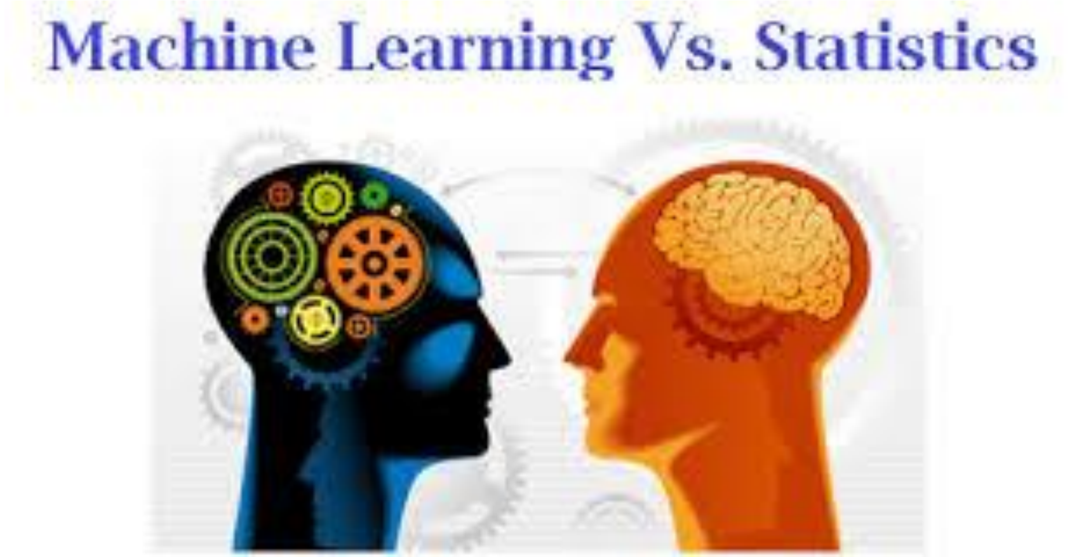
Life of a Statistician





New Life of a Statistician

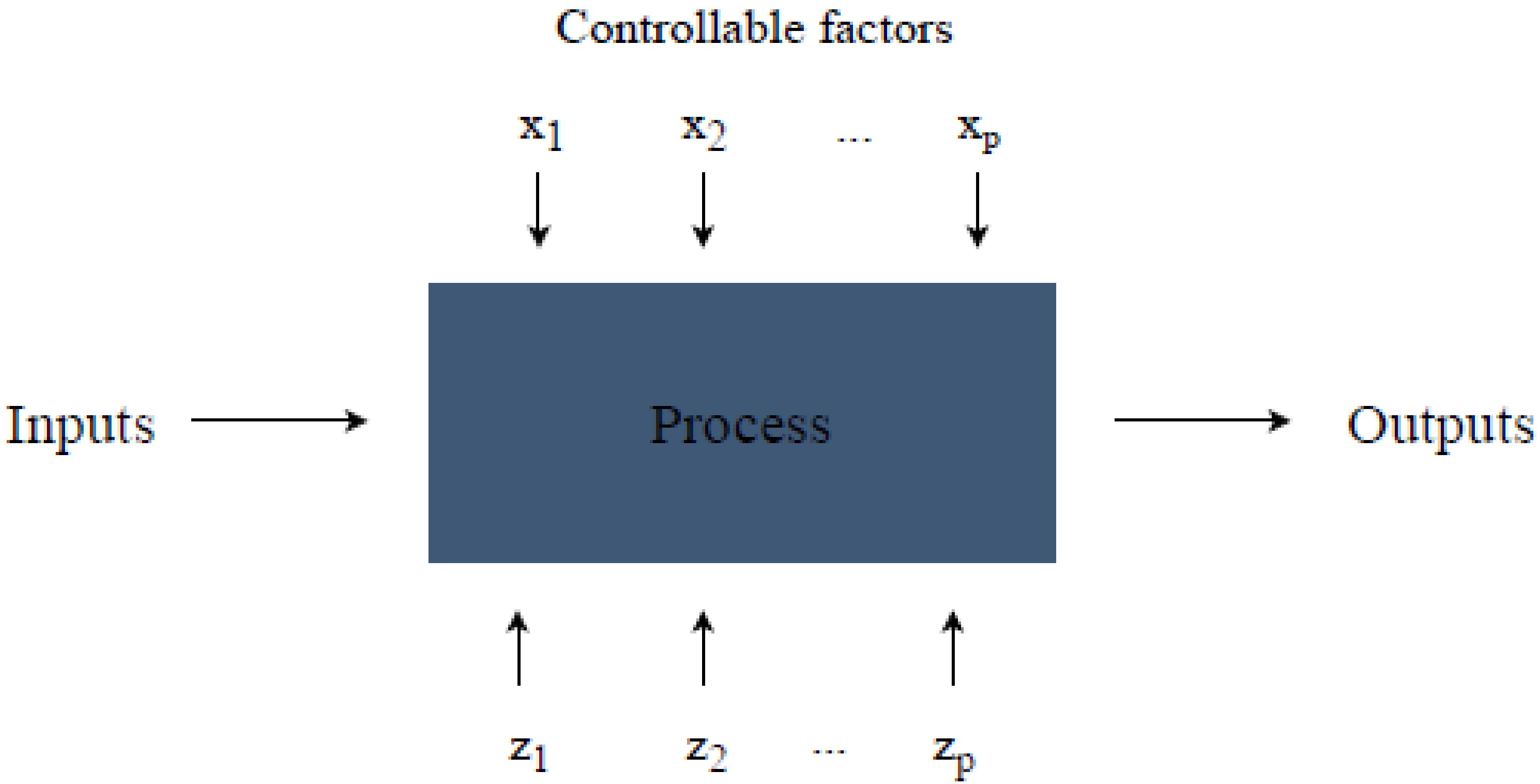
What now Industrial statistics?



- Those who advocates real understanding of what is happening in the relationships between variables and want to understand the causal effects and the science behind what is happening in the data.
- Those who are less concerned with the science and more concerned with getting good predictions. They seek less to understand causal effects.

Jensen 2020

A general model of an algorithm



Tuning hyperparameters in random forests

Table 5.3: Initial levels of the hyperparameters in the 16 run nonregular two-level design on the unsampled training set.

Factor	Hyperparameter	Low level (-)	High level (+)
A	*mtry*	2	6
B	*ntree*	250	750
C	*nodesize*	1	100
D	*cutoff*	(0.200,0.800)	(0.800,0.200)
E	*classwt*	(10,1)	(20,1)
F	*replace*	FALSE	TRUE

Table 5.4: Result of 16 runs with random forests on the unsampled training set with different values for hyperparameters *mtry*, *ntree*, *nodesize*, *cutoff*, *classwt* and *replace* using Tyssedal's design.

Run	*mtry*	*ntree*	*nodesize*	*cutoff*	*classwt*	*replace*	BACC
1	2	250	1	(0.200,0.800)	(20,1)	FALSE	0.500
2	6	250	1	(0.200,0.800)	(10,1)	FALSE	0.505
3	2	750	1	(0.200,0.800)	(10,1)	TRUE	0.500
4	6	750	1	(0.200,0.800)	(20,1)	FALSE	0.504
5	2	250	100	(0.200,0.800)	(10,1)	TRUE	0.500
6	6	250	100	(0.200,0.800)	(20,1)	TRUE	0.500
7	2	750	100	(0.200,0.800)	(20,1)	TRUE	0.500
8	6	750	100	(0.200,0.800)	(10,1)	FALSE	0.501
9	2	250	1	(0.800,0.200)	(10,1)	TRUE	0.651
10	6	250	1	(0.800,0.200)	(20,1)	TRUE	0.703
11	2	750	1	(0.800,0.200)	(20,1)	FALSE	0.626
12	6	750	1	(0.800,0.200)	(10,1)	TRUE	0.702
13	2	250	100	(0.800,0.200)	(20,1)	FALSE	0.503
14	6	250	100	(0.800,0.200)	(10,1)	FALSE	0.637
15	2	750	100	(0.800,0.200)	(10,1)	FALSE	0.577
16	6	750	100	(0.800,0.200)	(20,1)	TRUE	0.577

- Vatnedal (2020)

Volume 62 Number 1 February 2020

Technometrics number 1 2020

CONTENTS

Volume 62 Number 2 May 2020

ARTICLES

Technometrics number 2 2020

Industrial Statistics in 2020

• Dominating subjects	• A search on Oria for registration of books, journals, articles picture, music 1/1-19 - 19/9-20.
• DOE	• 655653
• SPC	• 480459
• Reliability	• 567875
• Machine /Statistical Learning	414444

What about the future?

- Old statistical problems are still there
- Hard to find any evidence of decline in the main subjects

New possibilities

- Comparing algorithms, tuning hyperparameters, combining DOE and Statistical learning.
- Improve data quality/Extract data with high quality (retrospective design)
- Good study design
- Monitoring predictions and features

Questions?

References

- Box, G., Hunter, W., and Hunter, J. (1978, 2005). Statistics for experimenters (1st and 2nd edition). John Wiley and Sons, NewYork.
- Box, G., and Tyssedal, J. (1996). Projective properties of certain orthogonal arrays. *Biometrika,* 83(4), pp. 950-955.
- Box G, Tyssedal J. Sixteen run designs of high projectivity for screening. *Communications in* Hussain, S., and Tyssedal, J. S. (2016). Projection Properties of Blocked Non-regular Two-level Designs. *Quality and Reliability Engineering International*, 32(8), pp 3011-3021.
- Hamre, Y. Preserving projection properties when regular two-level desigs are blocked.
- Jensen, W. A, (2020). Statistics=Analytics?. Quality Engineering, Vol 32(2), pp133-144.
- Jones-Farmer, L. A. (2019). Leveraging industrial statistics in the data revolution: The Youden Memorial Address at the 63rd Annual Fall Technical Conference.
- Quality Engineering, Vol 32 (2). pp 205-2011,
- Olsen,E. (1998). Statistisk analyse av en raffineringsprosess. Master Thesis NTNU.
- Steinberg, D. M. (2016). Industrial statistics: The challenges and the research. Quality Engineering, Vol 28(1), pp 45-59.
- Tyssedal, J. S., Grinde, H, Røstad, C. C. (2006) The Use of a 12-run Plackett-Burman Design in the Injection Moulding of a Technical Plastic Component. *Quality and Reliability Engineering International.* vol. 22 (6).
- Tyssedal, J. S., and Niemi, R. (2014). Graphical Aids for the Analysis of Two-Level Nonregular Designs. *Journal of Computational and Graphical Statistics,* 23(3).
- Vatnedal, R. P. (2020). Optimizing Predictive Performance of Random Forests by means of Design of Experiments and Resampling, with a Case Study in Credit Scoring.
- Wiik, E. H. (2014). Methods for Analyzing the 12 run Plackett-Burman Design. Master Thesis NTNU.