Design of Experiments in Ecological and Environmental Problems: methods and issues

Jorge Luis Romeu, Ph.D.
Research Professor, Syracuse University
Email: jlromeu@syr.edu
Web: http://www.linkedin.com/pub/jorge-luis-romeu/26/566/104
National Science Foundation, DC
May 28, 2013

Outline

- Problem statement
- Implementation problems
- Simulation Example
- Some applicable DOEs
- Other modeling alternatives
- Applications/extensions
- PASI in Latin America
- Conclusions

Problem statement

- Complexity of Environmental Problems
  - Too many variables in the system
  - Interactive/non linear structure
  - Difficulty in conducting experimentation
- Proposed solutions
  - Implement Design of Experiments (DOE)
  - In the Laboratory or with simulation models
  - EVOP approaches to experimentation

Examples of Environmental Projects

- Salinity, Ph., temperature, invasive species
  - In the survival of indigenous species
- Best mining and agricultural practices
  - In the life (length, quality) of specific species
- Contaminants, light, water velocity, flora
  - On indigenous species of the ecosystem
- Dam building and ecosystem destruction
- Difficulty to experiment in real environment
  - Or to re-create the complete environment in lab

A Recent NCER Announcement

Susceptibility and Variability in Human Response to Chemical Exposure
URL: http://www.epa.gov/ncer/rfa/2013/2013_star_chemical_exposure.html
Open Date: 06/10/2013 - Close Date: 09/10/2013

Summary: The U.S. Environmental Protection Agency (EPA), as part of its Science to Achieve Results (STAR) program, is seeking applications proposing research to study life stage and/or genetic susceptibility in order to better characterize sources of human variability in response to chemical exposure. The adverse outcome pathways (AOP) concept has the potential to serve as a framework for using susceptibility indicators, biomonitoring, and high throughput screening (HTS) data in an integrated manner to predict population responses to novel, potentially harmful, chemicals. While much emphasis has been placed on improved bio monitoring and HTS approaches, research is needed to understand the underlying factors that influence human susceptibility and to develop tools and methods for ID and use of susceptibility indicators in this context.

An Industrial Experiment Example

- Duress of bathroom tiles
  - Factors: time, temperature and concentration
  - Responses: average duress, variation
- Methods of experimentation
  - Lab: bake tiles in furnace at factor levels
  - Use actual tile manufacturers
    - In different places, that use different factors
- Problems associated with both approaches
  - Reproducing original conditions and inclusion

DOE Definition

- DoE consists in the planning activities for organizing and carrying out the “best” strategy for testing a statistical hypothesis
- Definition Keywords:
  - planning activities (before the event)
  - best strategy (seeks optimization)
  - hypothesis testing (statistical analysis)

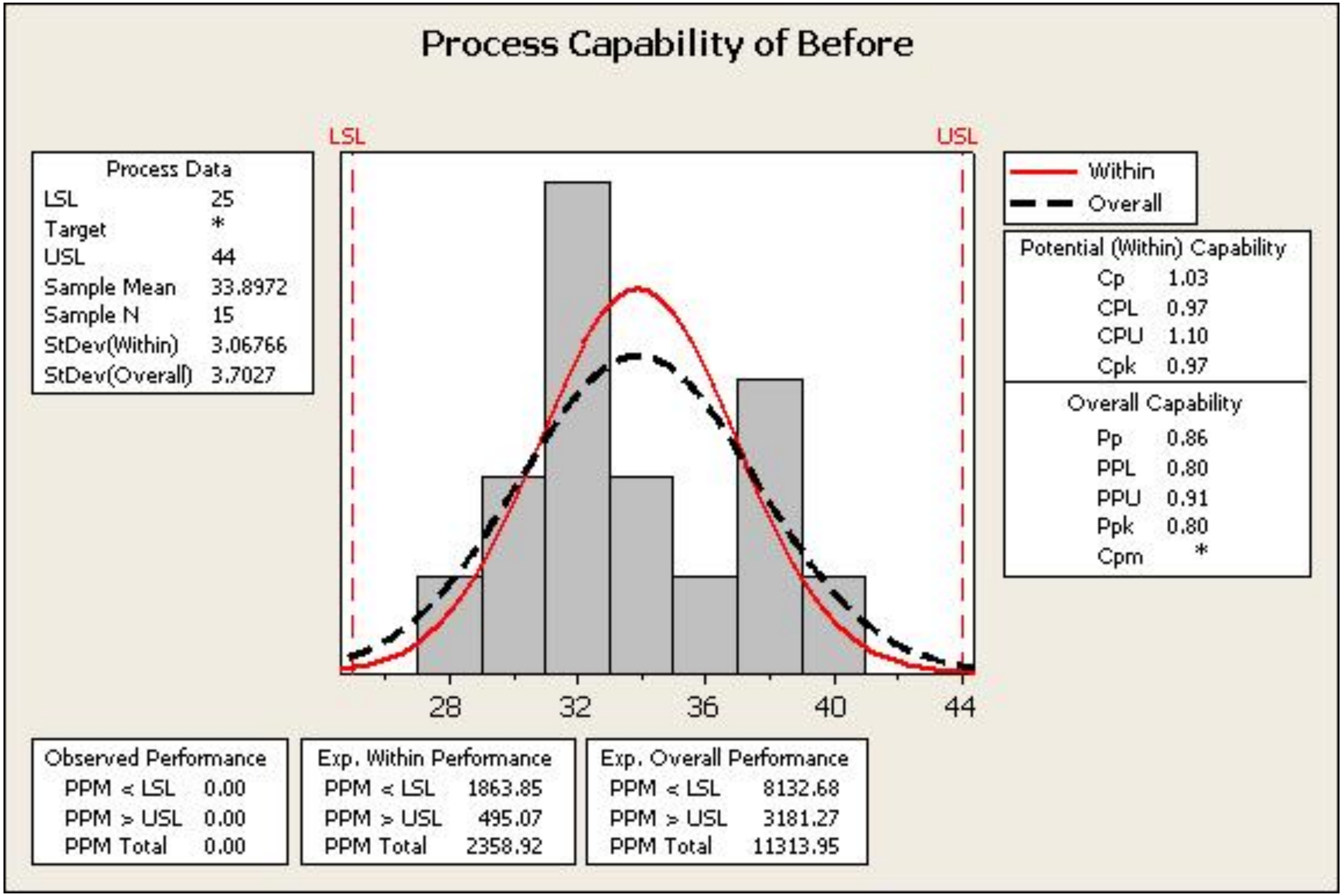
Steps to Perform DOE

- Set experimental objectives
- Select process variables
- Select an experimental design
- Execute the experimental design
- Check that data are consistent with experimental design assumptions
- Analyze and interpret results
- Conclude/Restart the loop

DOE Responses can be:

- Location parameter: average life length, number of individuals per unit, etc.
- Dispersion parameter: variance or standard deviation of life length, of individuals per unit.
- Certain factors impact variation, not location
- Variation has many useful applications
  - Comparison with upper/lower specification limits
- Variation is often slighted or ignored

Analyzing Variation as a Key Factor



Process Capability Indices:

$$C_p = \frac{U - L}{6\sigma}$$

$$C_{pk} = \frac{Min[U - \mu; \mu - L]}{3\sigma}$$

$$C_{pm} = \frac{U - L}{6\sqrt{\sigma^2 + (\mu - T)^2}}$$

$$C_{pmk} = \frac{Min[U - \mu; \mu - L]}{3\sqrt{\sigma^2 + (\mu - T)^2}}$$

Design of Experiments

- DoE considers several important issues:
  - desired precision of the results
  - significance level we can absorb
  - sample size required by problem
  - sampling schemes and estimators
- This requires the manipulation of the Factors
  - Before Experimentation Begins
  - Not always possible in environmental work

Planning a DoE Involves

- Determination of the response(s) Y
- Determination of the factors (X_1, X_2, X_3, X_4,…)
- Determination of the model functional form
- Determination of the interaction forms ($X_{1*}X_2$)
- Determination of the sample size (runs)
- Determination of the experimental precision
- Determination of the error we can absorb
- Determination of the randomization plan

Model Hypotheses are:

- Educated guesses
  - The result of experience or observation
- They are obtained by:
  - Restating the problem in statistical terms
- They are either true or false
  - The Null and Alternative hypotheses
  - Null (H_0): always the status quo
  - Alternative (H_1): negation of the Null!

Some Modeling Problems

- What if variances are different?
- Power of the test in experimental design
- Errors (α,β) provide the sample size
- Blocking when there are too many factors
- Assessing model assumptions (validation)
  - What happens with model violations?
  - How can we resolve such problems?
  - Not always done, or done incompletely

Choice of Sample Size

- Important Experimental Design Problem!
- Can be obtained by pre-specifying:
  - The precision of the experiment δ
  - Probabilities of types I and II errors (α,β)
  - Knowing the population variances σ^2
  - Obtain the required percentiles (z_α,z_β)
    - corresponding to the respective table values
    - for the respective probabilities $(1-\alpha)$ and $(1-\beta)$

Assessing Model Assumptions

- Data Independence
- Normality of the data
- Homogeneity of variances
- DOE Results are only valid
  - when all assumptions hold true
  - Check graphically, at the very minimum
- Robustness: degree of test validity under model assumption departures

Assumption Violations

- Lack of independence
- Heterogeneous variances
- Non-Normality of data
  - transformation of the data (Log, square root)
  - alternative non parametric procedures
- Always check model assumptions
  - At least graphically
  - to insure validity of your results!

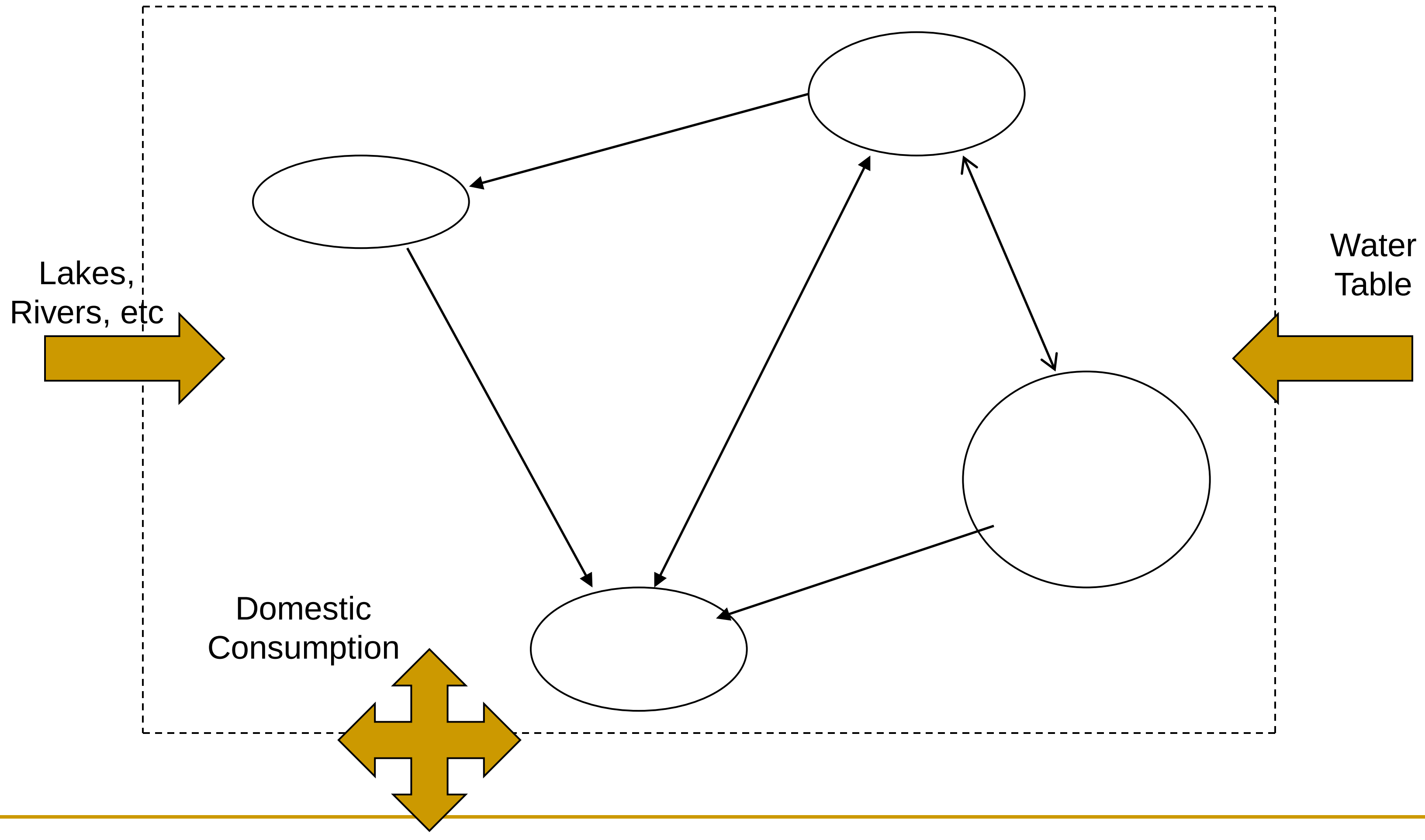
Three types of DOE experiments

- Laboratory Experiments
  - Not always possible to reproduce the situation
  - Certain elements may not be included
  - Missing factors and their interaction
  - That can also affect the response
- Simulation Experiments
  - Not always possible to model complete situation
- EVOP (Evolutionary operations)
  - Not entirely under experimenter's control

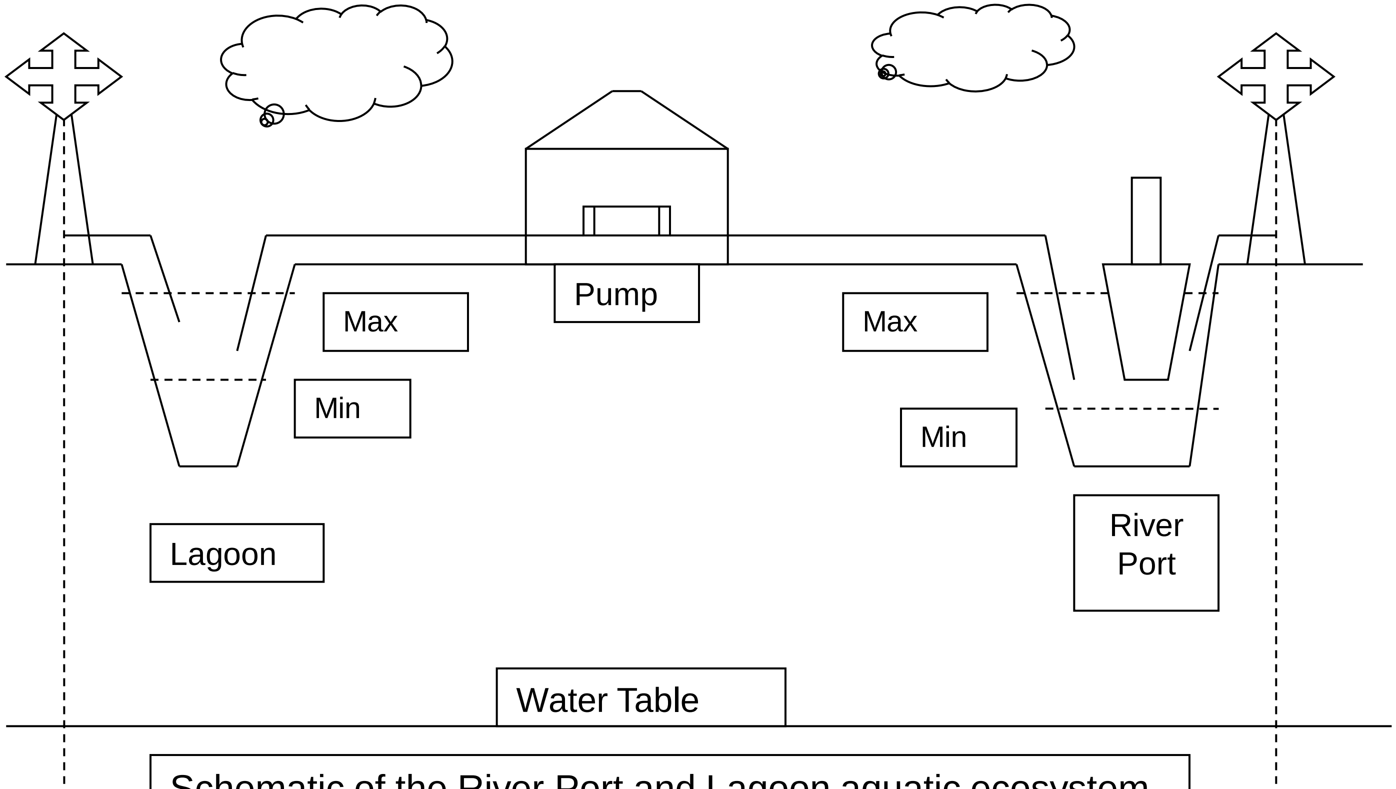
A Simulation Experiment Example

- Given a network of water masses
  - For both, civilian and industrial use
- Optimize some performance measures
  - e.g. operational, social, political, ecological
- Subject to a set of (conflicting) political, labor, socio-economic, etc. constraints
  - Maintaining levels of production, employment
  - Tax revenues, social services, economic, etc.

A Network of Interconnected Water Masses



Example: River Port w/Lagoon



Schematic of the River Port and Lagoon aquatic ecosystem.

Controlled Variables: Economic

- Replenishing Levels (MIN)
- Reservoir Capacity (MAX)
- Ordering Schedule
- Transfer Policy
- Usage Policy
- Shortage Policy
- Profitability
- System's Initial Conditions

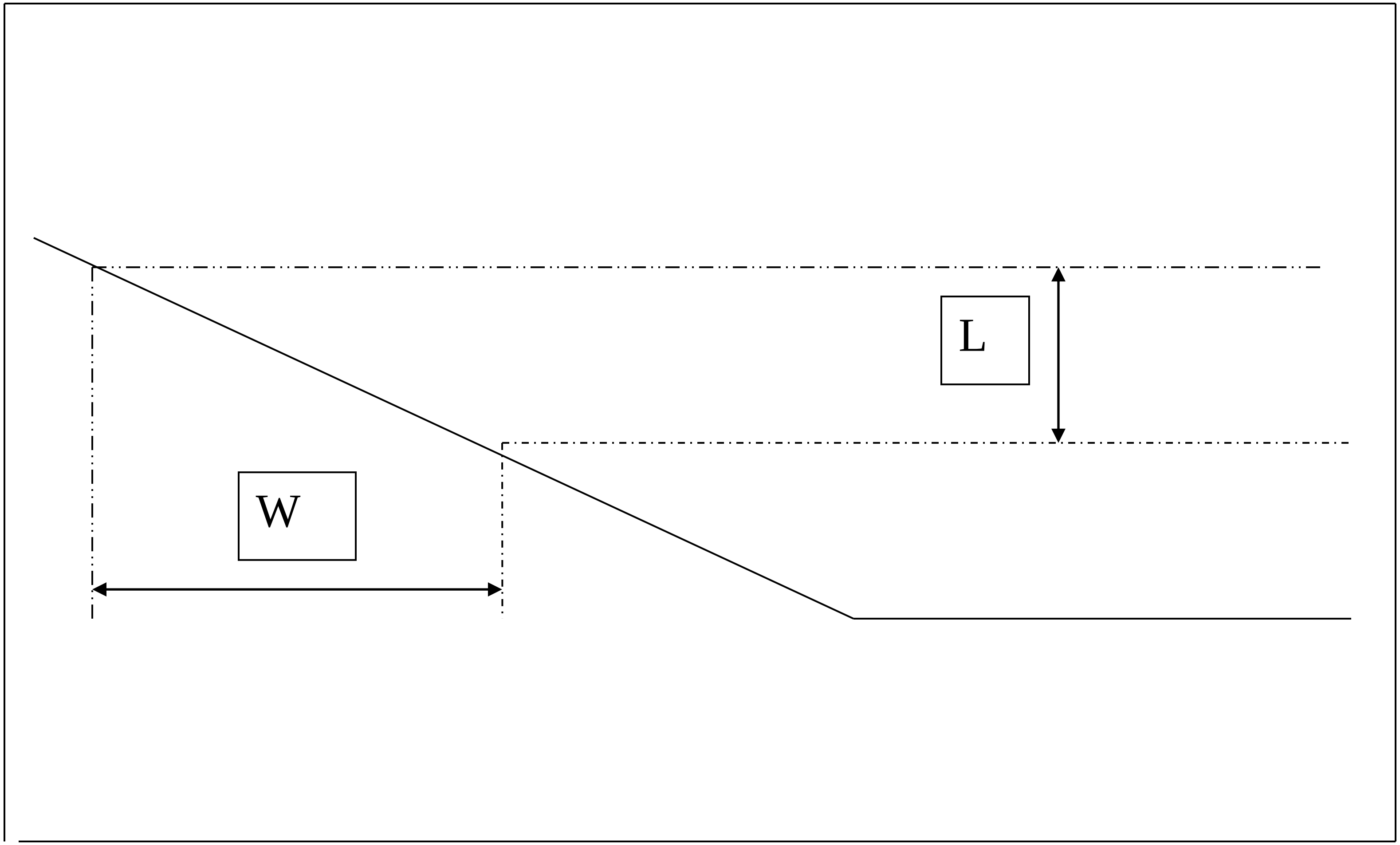
Controlled Variables: Social

- Allocation to each sector
- Size of the Reservoirs
- Transfer Policy
- Generation of electricity
- Hospitals and schools
- Transportation uses
- Recreation uses

Controlled Variables: Ecologic

- Wetland Area
- Wetland Depth
- Transfer Speed
- Water Table Use
- Pollution Level
- Fish/Foul Population

WetLand v. Level



Uncontrolled Variables

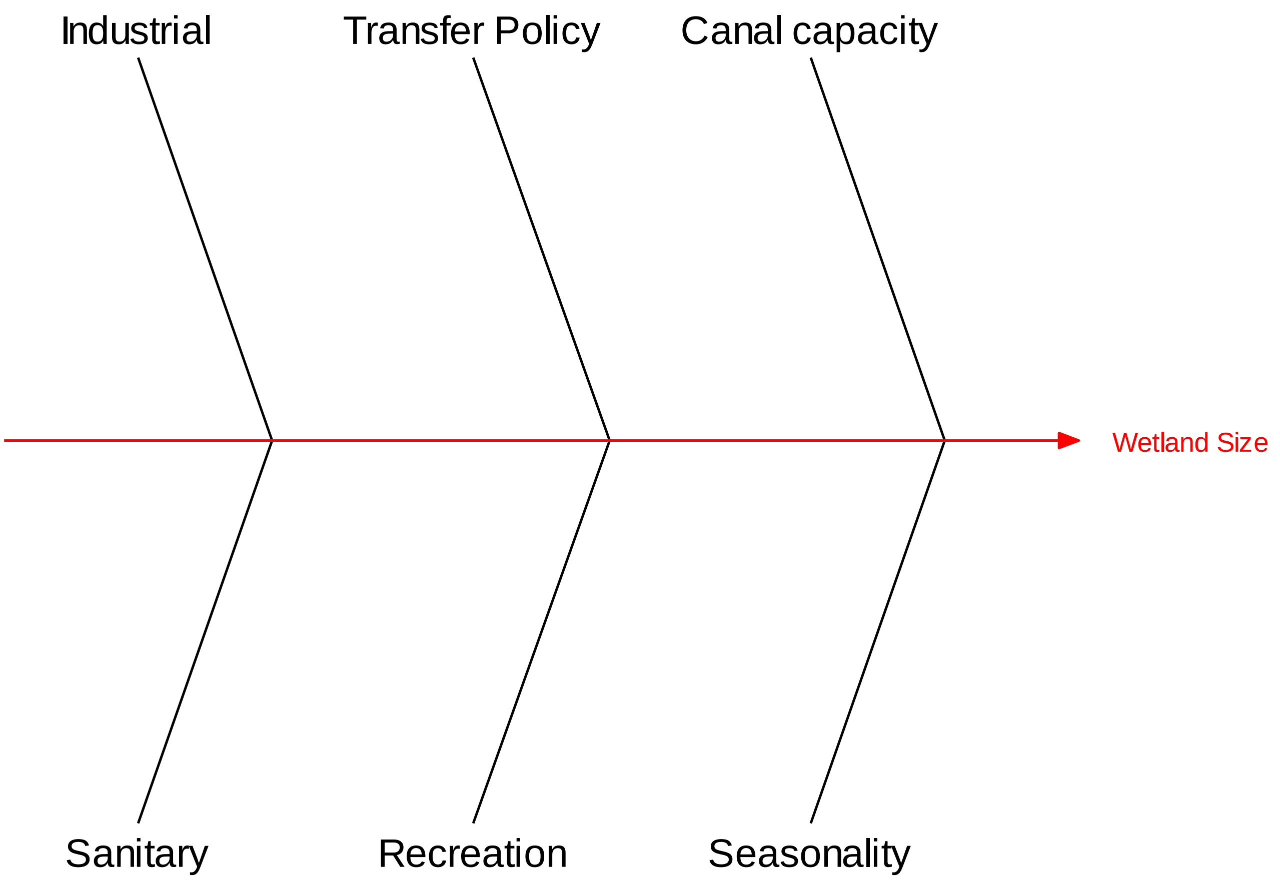
- ECONOMIC
- Political issues
- Labor issues
- Water Theft
- Water Leaks
- Markets
- Financial

- ECOLOGIC
- Evaporation
- Temperature
- Salinity
- Reproduction
- Weather
- Water Table

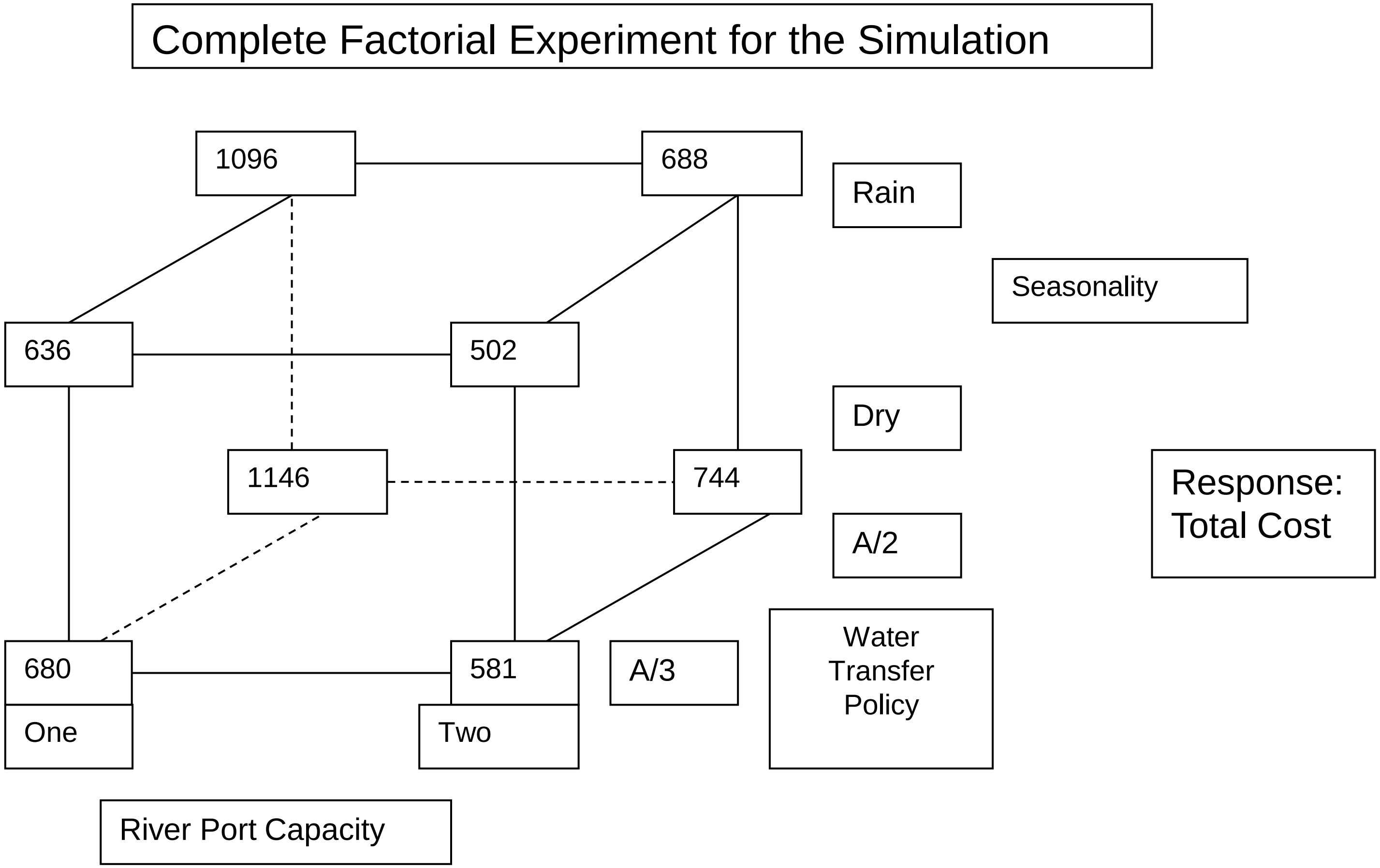
And Associated Costs

- Of Importing Water from other places
- Transferring from Social to Economic
- Allocation to various constituencies
- Of Water shortages and rationing
- Indirect costs (labor, political, social)
- Ecological costs (degradation, loss)
- Total costs (compound response)

Simlation of Finger Lakes Ecosystem

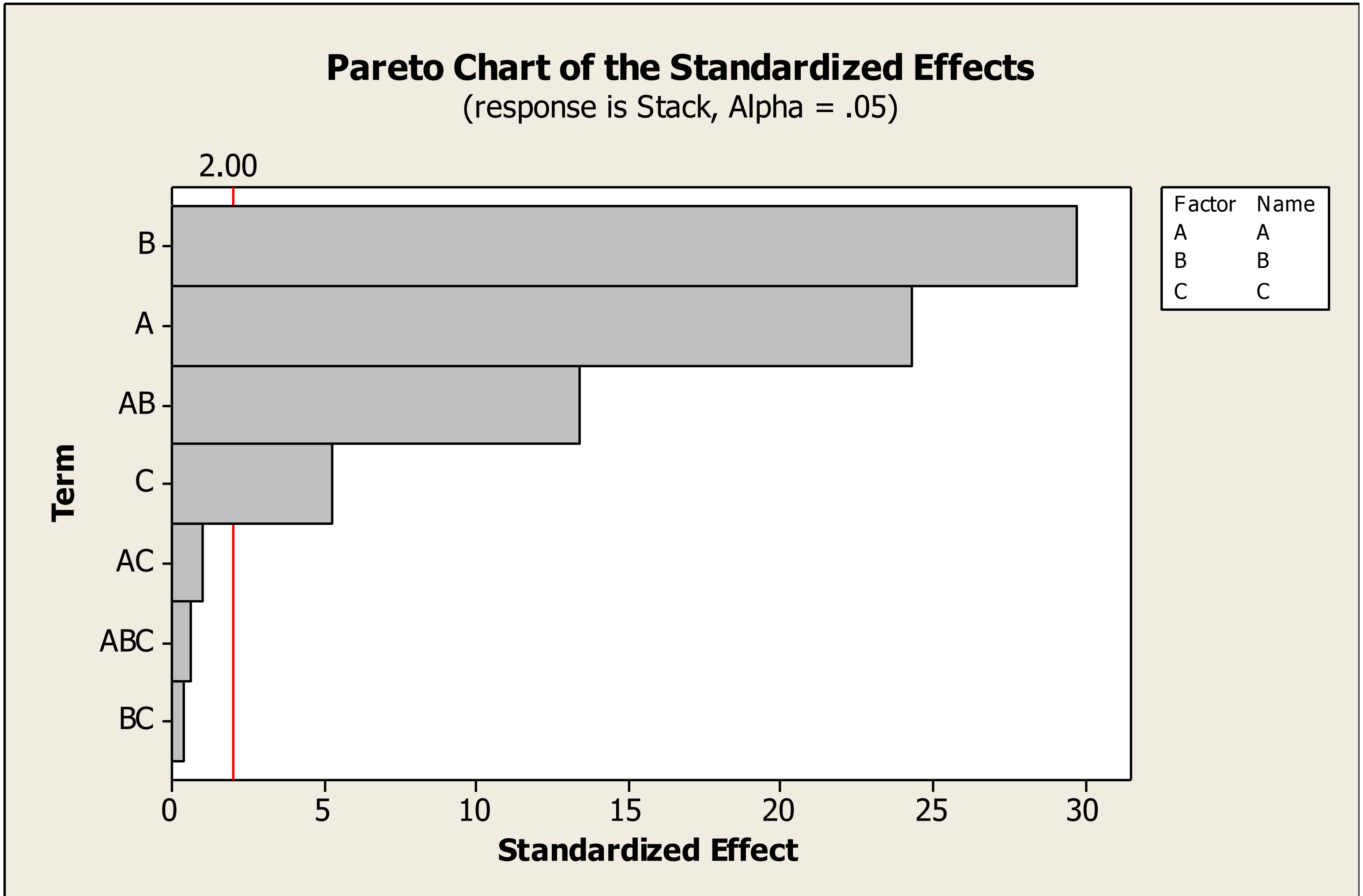


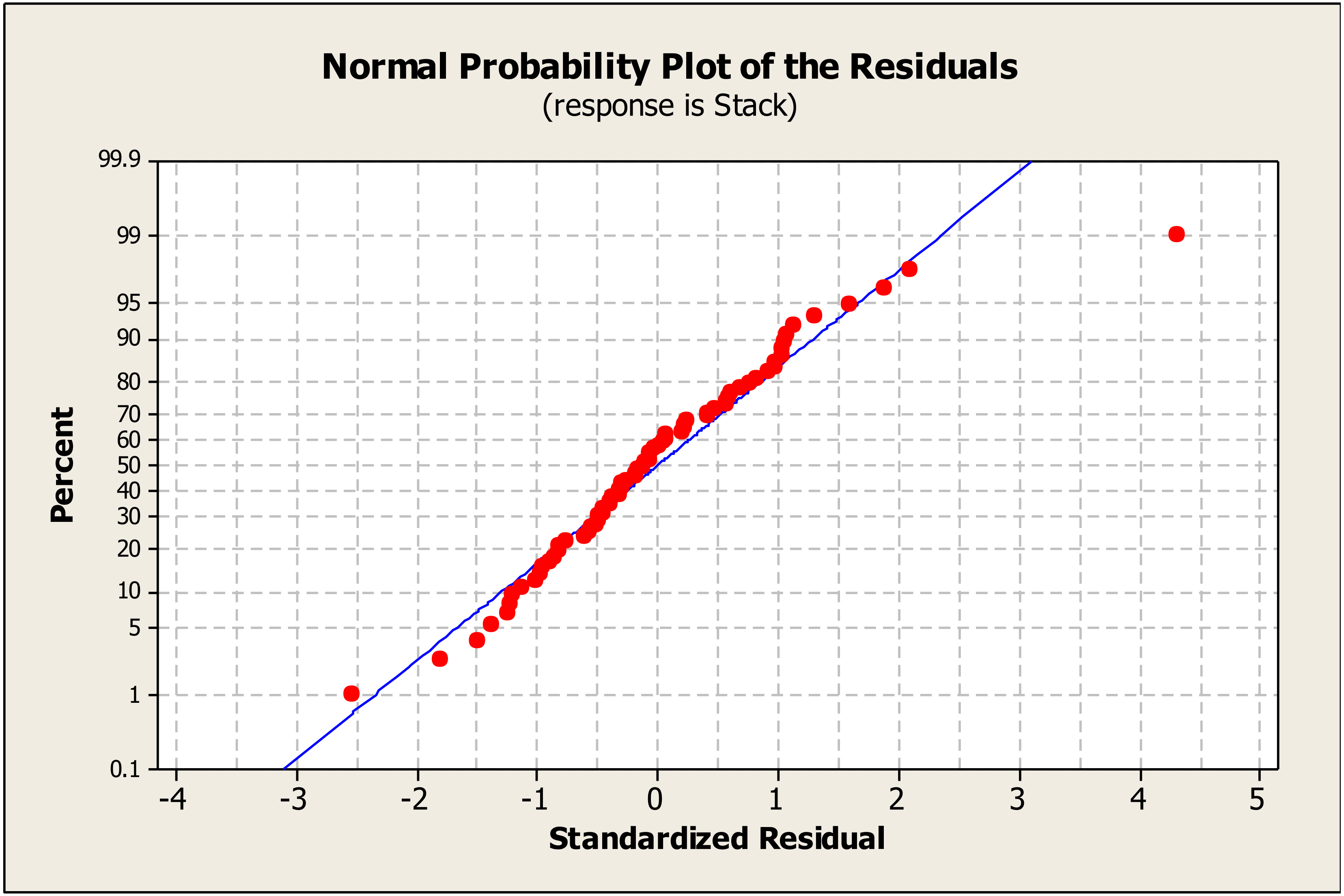
Example of a Simple DOE



Experimental Results

- Factor 1: Ecosystem capacity
  - of the Lake
  - of the River Canal
- Factor 2: Water Transfer Policy
  - between water masses and Water Table
- Factor 3: Seasonality (Spring/Fall)
- Interaction: F1 * F3
- All other variables were non-significant


Pareto Chart of the Standardized Effects
(response is Stack, Alpha = .05)
2.00
B
A
AB
C
AC
ABC
BC
Term
0
5
10
15
20
25
30
Standardized Effect
Factor Name
A A
B B
C C



Normal Probability Plot of the Residuals
(response is Stack)
99.9
99
95
90
80
70
60
50
40
30
20
10
5
1
0.1
Percent
-4
-3
-2
-1
0
1
2
3
4
5
Standardized Residual

Statistical Results

Table 2: Analysis of Variance Table for the Simulation Experiment

Source	D. F.	Mean Square	F Value	P-Value
River Canal Capacity	1	1219401	588.71	0.000
Water Transfer Policy	1	1828892	882.96	0.000
Seasonality (Spring/Fall)	1	58186	28.09	0.000
Enviro-Site x Policy	1	373104	180.33	0.000

Table 3: Examples of model-derived quantitative information

Factor	Change Effected	Effect on Response
River Canal Capacity	One to two ships capacity	Size decreases
Water Transfer Policy	Transfer: 1/3 to 1/2 of water mass availability	Size increase:
Seasonality	Spring into Fall Season	Size decreases
Capacity x Seasonality	Spring/one ship; Fall/Two	Size decreases:

Some Modeling Applications

- Design and Optimization of Systems
- Identification of System Key Factors
- Analysis of System Key Factors
- Arbitration and Conflict Resolution
- Evaluation of Decisions/Strategies
- Evaluation of Robust Strategies
- Trade-offs and Sensitivity Analyses
- What-if, Time to catastrophic fails, etc.

Composite Objective Functions

Ecologic: Xi is number of occurrences of ith item:

$$f(x_1, \cdots, x_p) = \sum_i v_i x_i ;.with: \sum v_i = 1$$

Economic: Yi = aiXi is cost of No. ith item occurrences:

$$g(x_1, \cdots, x_p) = \sum_i \lambda_i y_i ;.with: \sum \lambda_i = 1$$

$$l(w_1, \cdots, w_n) = \sum_i \delta_i w_i ;.with: \sum w_i = 1$$

Arbitration and Trade-Off: α is the preference or weight:

$$H(g, l) = \alpha g + (1 - \alpha) l ;.with: 0 < \alpha < 1$$

Example of modeling approach:

- Minimize Total Water Operations Cost
- Subject to:
  - Maintaining specified labor levels
  - Reducing pollution to specified levels
  - Maintaining specified social levels
  - Maintaining specified consumption levels
  - Increasing overall health indices

Trade-Off Examples

Scenario	Ecologic	Health	Industry	Education	Recreation	Other
Best Ecologic	**X1**	**Y1**	**Z1**	**W1**	**L1**	**M1**
Best Health	**X2**	**Y2**	**Z2**	**W2**	**L2**	**M2**
Best Industry	**X3**					
Best Education	**X4**					
Best Recreation	**X5**					
Best Other	**X6**					

Analyze Maxi-min and Mini-max results

Some DOE Model Limitations:

- Analyzes limited variables (here, k=3)
  - For, 2^K Factors/Interacts are generated
- The Effect of Interaction, when k > 2
  - Can affect results, if present and strong
- Need to find Robust Responses
  - Handling specific “noise variables”
- Need to Identify “significant few” variables
  - To reduce model Size, maintaining Info level.

Consequences … and Solutions

- Large number of factors to analyze
  - Strong factor interaction may exist
  - Dependent on the model structure
  - Requires special methods for analysis
- Different objective of models derived:
  - To describe/study, forecast or control
- Robust Parameter analysis capability
  - To derive a response equation that is
  - Resilient to “noise” or uncontrolled factors

Some Variable Id Methods

- Full Factorial Designs
- Fractional Factorial Designs
- Plackett-Burnam Designs
- Latin Hypercube Sampling
  - Regression Selection methods
  - Principal Components/PCA
- Other modeling approaches:
  - Taguchi Methodology

Full Factorials

- Most expensive (in time and effort)
  - Prohibitive with current number of factors
- Most comprehensive information
  - Provides info on all factor interactions
- Two Examples with a 2^3 Full Factorial
  - First case: mild interaction (AB only)
  - Second: strong and complex interaction
  - Notice how the Model-Estimations vary

Example 2^3 Full Factorial Design:

- **Variables Used**
- A = Replenishing Levels (MIN)
- B = Reservoir Capacity (MAX)
- C = Transfer Policy
- Mild interaction assumed
- A* B only

Full Factorial Experiment 2^3

Run	A	B	C	AB	AC	BC	ABC	Avg.
1	-1	-1	-1	1	1	1	-1	-1.07
2	1	-1	-1	-1	-1	1	1	3.72
3	-1	1	-1	-1	1	-1	1	-0.58
4	1	1	-1	1	-1	-1	-1	12.04
5	-1	-1	1	1	-1	-1	1	7.75
6	1	-1	1	-1	1	-1	-1	15.45
7	-1	1	1	-1	-1	1	-1	11.09
8	1	1	1	1	1	1	1	18.31
TotSum								66.71
Effect	8.08	3.75	9.62	1.84	-0.62	-0.65	-2.08	

Regression Estimations

RegCoef	A	B	C	AB	b0
Estimat.	4.04	1.88	4.81	0.92	8.34
TRUE	4	2	5	1	10

Meta Model: $Y_{ijkl} = 8.33 + 4.04A + 1.88B + 4.81C + 0.92AB$

True Model: $Y = 10 + 4*A + 2*B + 5*C + AB + \varepsilon$

Mild Interaction (AB only)

Fractional Factorial Designs

- Analyzes only a Fraction of Full Factorial
  - Reduces substantially time/effort
  - Confounding of Main Effects/Interactions
  - If Interactions present, this is a problem
  - Only for Powers of Two (no. of runs)
- Numerical Example: Half Fractions
  - Of the previous Full Factorial –and others
  - Assess Model-Estimation agreement

Fractional Factorials

First Fraction: L1

Run	**A**	**B**	**C=AB**	**Avg.**
1	1	-1	-1	-0.33
2	-1	1	-1	-0.33
3	-1	-1	1	-0.33
4	1	1	1	1.00
TotSum				0.00
Effect	7.429	3.130	11.460	
Signif.	No	No	Yes	

$$Y_1 = 7.3 + 3.71A + 1.57B + 5.73C^*$$

C *: Factor C is confounded with AB

Second Fraction: L2

Run	**A**	**B**	**C=AB**	**Avg.**
1	-1	-1	-1	-1.00
2	1	1	-1	0.33
3	1	-1	1	0.33
4	-1	1	1	0.33
TotSum				0.00
Effect	8.728	4.375	7.784	
Signif.	Yes	No	Yes	

$$Y_2 = 8.33 + 4.36A + 2.18B + 3.89C^*$$

Untangling Confounded Structure

(L1+L2)/2	**8.079**	**3.753**	**9.622**
(L1-L2)/2	-0.649	-0.623	1.838
Effects	**8**	**4**	**10**

Notice how, by averaging both Half Fraction results, we obtain the Full Factorial results again.

True Model: $Y = 10 + 4{*}A + 2{*}B + 5{*}C + AB + \varepsilon$

Re-analyzing the 2^3 Full Factorial: The same Variables are used, but With Stronger Interaction

- A = Replenishing Levels (MIN)
- B = Reservoir Capacity (MAX)
  - C = Transfer Policy
- Stronger interaction assumed
  - A*B, A*C, B*C
  - Overall: A*B*C

Full Factorial: Complex, Stronger Interaction

	Model Parameters						
Variables	**A**	**B**	**C**	**AB**	**AC**	**BC**	**ABC**
RegCoef	**3**	**-5**	**1**	**-12**	**8**	**-10**	**-15**
RegEstim	1.94	-4.38	1.73	-12.14	7.34	-10.52	-15.26
MainEffEst	3.88	-8.76	3.47	-24.28	14.68	-21.05	-30.51
MainEffcts	**6**	**-10**	**2**	**-24**	**16**	**-20**	**-30**

Var. of Model	12.5173	**StdDv**	3.53799
Var. of Effect	2.0862	**StdDv**	1.44437
Student T (0.025DF)	2.47287		
C.I. Half Width	3.57177		

Factor	**A**	**B**	**C**	**AB**	**AC**	**BC**	**ABC**
Signific.	Yes	Yes	No	Yes	Yes	Yes	Yes

True Model and Estimated Meta Model:

Y = 3A - 5B + C - 12AB + 8AC - 10BC - 15ABC

RegEstim 1.94A -4.38B 1.73C -12.14AB +7.34AC -10.52BC -15.26ABC

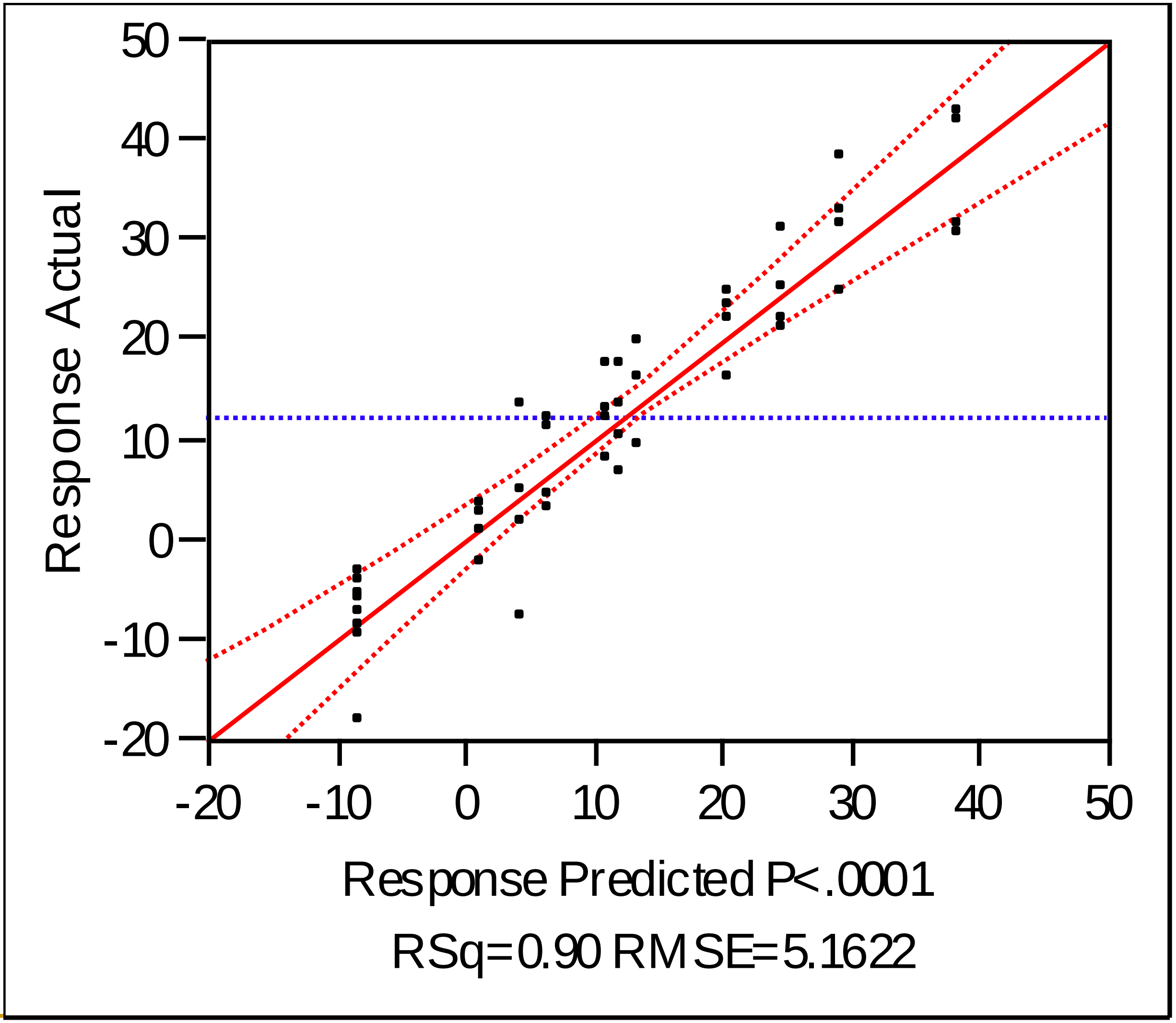
Corresponding Half Fractions

		Half	**Fraction**	**Analysis:**					
			First	**Half(a)**					
Run	**A**	**B**	**C=AB**	**Y1**	**Y2**	**Y3**	**Avg.**	**Var**	**Model**
2	1	-1	-1	-15.03	-16.54	-16.04	-15.87	0.59	-14
3	-1	1	-1	7.18	9.21	5.28	7.22	3.87	6
5	-1	-1	1	-16.75	-19.75	-22.02	-19.51	6.97	-22
8	1	1	1	-31.61	-27.62	-33.04	-30.76	7.89	-30
TotSum				-56.21	-54.7	-65.82	-58.91	19.32	
Effect	-17.17	5.92	-20.81		ModlVar.	4.83	StdDev=	2.2	EffVar
Signif.	Yes	Yes	Yes		T(.975,df)	2.75	CI-HW=	3.49	StdDev

			Second	**Half(b)**					
Run	**A**	**B**	**C=-AB**	**Y1**	**Y2**	**Y3**	**Avg.**	**Var**	**Model**
1	-1	-1	-1	-5.64	-0.28	9.43	1.17	58.32	2
4	1	1	-1	4	1.47	2.49	2.65	1.62	2
6	1	-1	1	49.73	54.94	56.86	53.84	13.62	54
7	-1	1	1	5.99	7.88	2.56	5.48	7.26	2
TotSum				54.08	64.01	71.34	63.14	80.82	
Effect	24.92	-23.44	27.75		ModlVar.	20.2	StdDev=	4.49	EffVar
Signif.	Yes	Yes	Yes		T(.975,df)	2.75	CI-HW=	7.14	StdDev
(a+b)/2	3.88	-8.76	3.47	**MainEff**	**"C"**				
(a-b)/2	-21.05	14.68	-24.28	**Interact**	**C=AB**				
Coefs	6	-10	2						

NOTE: FRACTIONAL FACTORIAL RESULTS, GIVEN THE STRONG INTERACTIONS, ARE POOR.

Plot Real vs. Prediccion



Plackett-Burnam (PB) Designs

- Are Fractional Factorial (FF) DOEs
- Analyses “holes” between adjacent FFs
- Reduces time/effort, considerably
- Confounding of Main Effects/Interactions
- Numerical Example: 11 main effects
  - Compare PB to a 2^11 Full Factorial
  - Not all Interactions are strong/significant
- Counter Example: strong interactions

Plackett-Burnam w/o Interaction

- A=Replenishing Levels (MIN)
- B=Reservoir Capacity (MAX)
- C=Ordering Schedule
- D=Transfer Policy
- E=Allocation to each sector
- F=Size of the Reservoirs
- G=Generation of electricity
- H=Hospitals and schools
- I=Wetland size
- J=Water Table
- K=Fish/Foul Population

Placket-Burnam Design (no interaction)

Run	A	B	C	D	E	F	G	H	I	J	K	**Avg**
1	1	-1	1	-1	-1	-1	1	1	1	-1	1	36.14
2	1	1	-1	1	-1	-1	-1	1	1	1	-1	24.39
3	-1	1	1	-1	1	-1	-1	-1	1	1	1	0.5
4	1	-1	1	1	-1	1	-1	-1	-1	1	1	-5.96
5	1	1	-1	1	1	-1	1	-1	-1	-1	1	2.62
6	1	1	**1**	-1	1	**1**	-1	1	-1	-1	-1	31.26
7	-1	1	1	1	-1	1	1	-1	1	-1	-1	21.12
8	-1	-1	1	1	1	-1	1	1	-1	1	-1	-10.54
9	-1	-1	-1	1	1	1	-1	1	1	-1	1	15.92
10	1	-1	-1	-1	1	1	1	-1	1	1	-1	12.02
11	-1	1	-1	-1	-1	1	1	1	-1	1	1	7.33
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	11.66
Factors	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**Bo**
RegCoef	**6**	**2**	**0**	**-4**	**-6**	**0**	**-2**	**4**	**8**	**-8**	**0**	**12**
RegEst.	**4.5**	**2.3**	**-0.1**	**-4.3**	**-3.6**	**1.4**	**-0.8**	**5.2**	**6.1**	**-7.6**	**-2.8**	**12.2**
MainEff	12	4	0	-8	-12	0	-4	8	16	-16	0	n/a
EstimEff	9.1	4.7	-0.2	-8.6	-7.2	2.8	-1.5	10.4	12.3	-15.2	-5.6	12.2
Signific.	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes

Plackett-Burnam with Strong Interaction

- A=Replenishing Levels (MIN)
- B=Reservoir Capacity (MAX)
- C=Ordering Schedule
- D=Transfer Policy
- E=Allocation to each sector
- F=Size of the Reservoirs
- G=Generation of electricity
- H=Hospitals and schools
- I=Wetland size
- J=Water Table
- K=Fish/Foul Population

Model with Strong Interaction Structure

Factors	A	B	C	D	E	F	G	H	I	J	K	Bo
RegCoef	6	2	0	-4	-6	0	-2	4	8	-8	0	12

Interaction: **2*A*B-4*H*I+G*J+D*E**

Plackett-Burnam (n=12 rows) Analysis Results:

Factors	A	B	C	D	E	F	G	H	I	J	K
MainEff	12	4	0	-8	-12	0	-4	8	16	-16	0
FacEstim	-98.6	61.1	41.3	-86.5	98.4	66.4	79.7	51.8	-26.6	37.6	-96.0
RegPar.	6	2	0	-4	-6	0	-2	4	8	-8	0
RegEstim	-49.3	30.5	20.6	-43.2	49.2	33.2	39.8	25.9	-13.3	18.8	-48.0
Signific.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Results are seriously confounded and numerically erroneous.

There are two groups of significant variables after Plackett-Burnam:

Positive: B, C, E, F, G, H, J;

and

Negative: A, D, I, K.

We Perform a Resolution IV FF

To one of the two groups

Re-Analyzing the Group of Positive Variables: B, C, E, F, G, H, and J

- B=Reservoir Capacity (MAX)
- C=Ordering Schedule
- E=Allocation to each sector
- F=Size of the Reservoirs
- G=Generation of electricity
- H=Hospitals and schools
- J=Water Table

Performing a Resolution IV FF to the "Positive" group: B, C, E, F, G, H, J

Factors	B	C	E	F	G	H	J	Bo
TRUE	**12**	**4**	**0**	**-8**	**-12**	**0**	**-4**	**12**
EffectEstim	12.14	2.53	1.17	-7.20	-11.82	0.39	-3.49	13.59
RegCoef	6	2	0	-4	-6	0	-2	12
RegEst.	6.07	1.26	0.59	-3.60	-5.91	0.19	-1.75	6.80
Signific.	Yes	Yes	No	Yes	Yes	No	Yes	

Notice how, once all the (erroneously estimated) variables of the "same sign" were re-analyzed as a sub-group. Plackett-Burnam estimations then became closer to the True parameter values, both in sign and in magnitude.

Latin Hypercube Example

Assume we have a three dimensional (p = 3) problem in variables B, I, J (reservoir capacity; wetland size and water table use) and that these are respectively distributed Normal, Uniform and Exponential,. Assume that we want to draw a random sample of size n = 10. Divide each variable, according to its probability distribution, into ten equi-probable segments (Prob. = 0.1 = 1/10), identifying each segment with integers 1 through 10. Then, draw a random variate (r.v.) from each of the ten segments, for each of the three variables B, I, J. Finally, obtain the 10! permutations of integers 1 through 10. Randomly assign one of such permutations (e.g. segments 2,1,5,4,6,9,8,10,7 for B), to each of the variables, select the corresponding segment r.v., and form the vector sample, as below:

Example of Latin Hypercube Sampling Segments

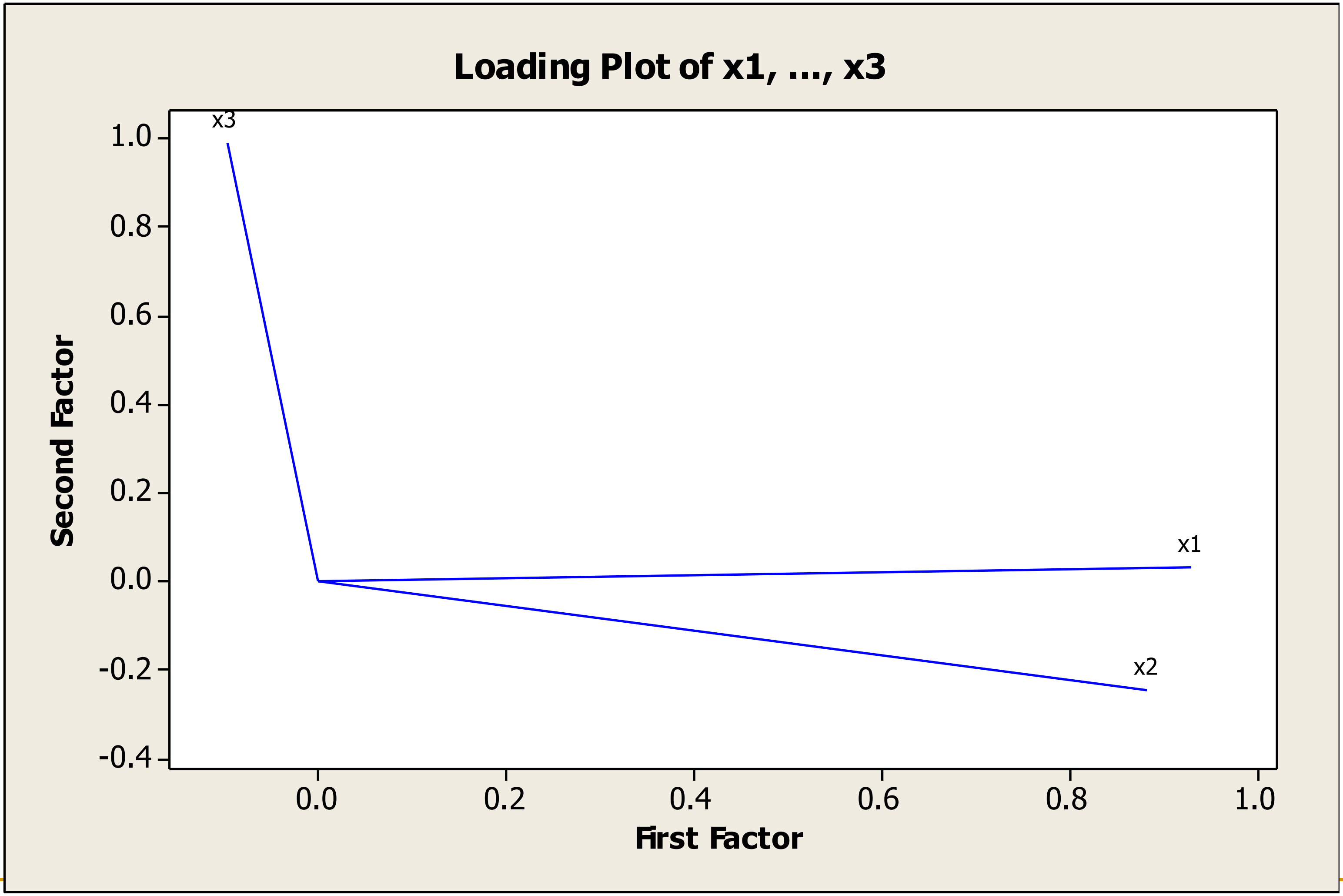
Sample	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**10th**
B	2	3	1	5	4	6	9	8	10	7
I	4	2	7	1	5	9	10	8	6	3
J	8	6	2	7	1	5	4	3	9	10

Latin Hypercube Sampling Modeling

- Multiple regression analysis approach
  - Sampling at “best” points in sample space
- Regression selection methods
  - To obtain most efficient Meta Model set
- Provides a list of Alternative Meta Models
  - Some, not as efficient -but close enough
  - Their factors can be “controlled” by the user
- Can be reduced via Principal Components

Example of Varimax Factor Rotation :
Project Variables X1 and X2 on F1
Then, Project Variable X3 on Factor 2.

Variable	Factor1	Factor2
x1	0.930	0.030
x2	0.883	-0.249
x3	-0.097	0.989



Taguchi Methodology

- Analyzes both Location and Variation
  - Of the performance measure of interest
- Best combination of both these together
  - To obtain the most efficient Model
- Optimize Location, resilient to Variation
- Minimize Variation, resilient to Location
- Determine regions of joint optimality
- Determine Variation is NOT an issue
- Done equivalently by implementing a DOE.

Taguchi SN Ratios

The preferred parameter settings are determined through analysis of the "signal-to-noise" (SN) ratio, where factor levels that maximize appropriate SN ratio are optimal. There are three standard types of SN ratios that depend on the desired performance response:

- Smaller the better (for making system response as small as possible):

$$SN_S = -10 * \text{Log}[1/n\ (\sum y_i^2)]$$

- Nominal the best (for reducing variability around a target):

$$SN_T = 10 * \text{Log}\ (y^2 / s^2)$$

- Larger the better (for making system response as large as possible):

$$SN_S = -10 * \text{Log}[1/n\ (\sum 1/y_i^2)]$$

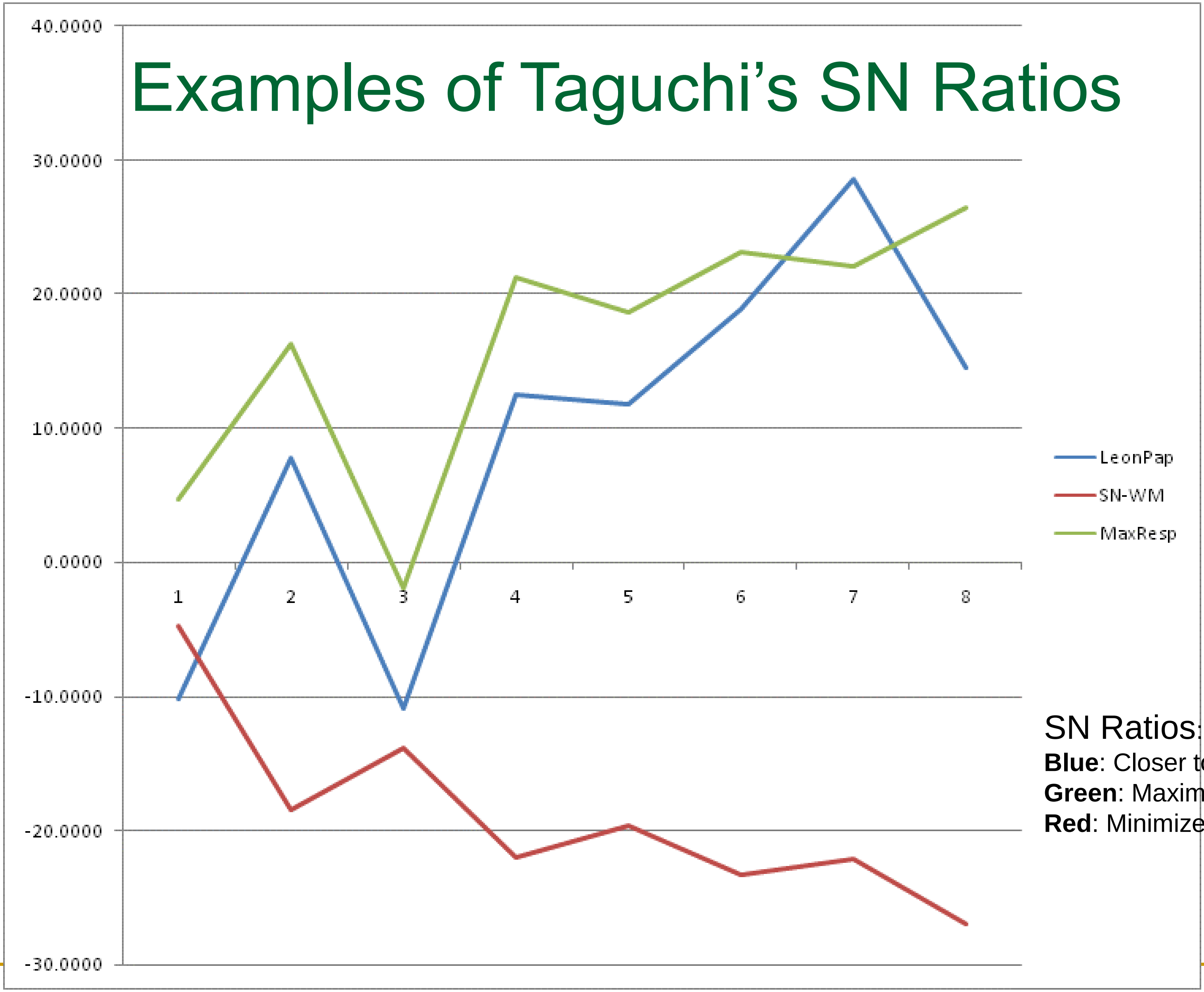
These SN ratios are derived from the quadratic loss function.

Example of Taguchi Methodology

X1	X2	X3	X4	X5	1	2	3	4	Var	LnVar	Average	TaguchiSN
1	1	1	-1	-1	194	197	193	275	1616.25	7.39	214.75	-46.75
1	1	-1	1	1	136	136	132	136	4.00	1.39	135.00	-42.61
1	-1	1	-1	1	185	261	264	264	1523.00	7.33	243.50	-47.81
1	-1	-1	1	-1	47	125	127	42	2218.92	7.70	85.25	-39.51
-1	1	1	1	-1	295	216	204	293	2376.67	7.77	252.00	-48.15
-1	1	-1	-1	1	234	159	231	157	1852.25	7.52	195.25	-45.97
-1	-1	1	1	1	328	326	247	322	1540.25	7.34	305.75	-49.76
-1	-1	-1	-1	-1	186	187	105	104	2241.67	7.71	145.50	-43.59

- **VARIABLES ANALYZED**
- Response: Wet Land Size
- X1=Reservoir Capacity (MAX)
- X2=Generation of electricity
- X3=Hospital Capacity
- X4=Social Services
- X5=Fish/Foul Population
- Z1 and Z2 are two ***noise*** variables

Examples of Taguchi's SN Ratios



SN Ratios:

Blue: Closer to Target

Green: Maximize Yield

Red: Minimize Yield

Analysis for Joint Location-Variance

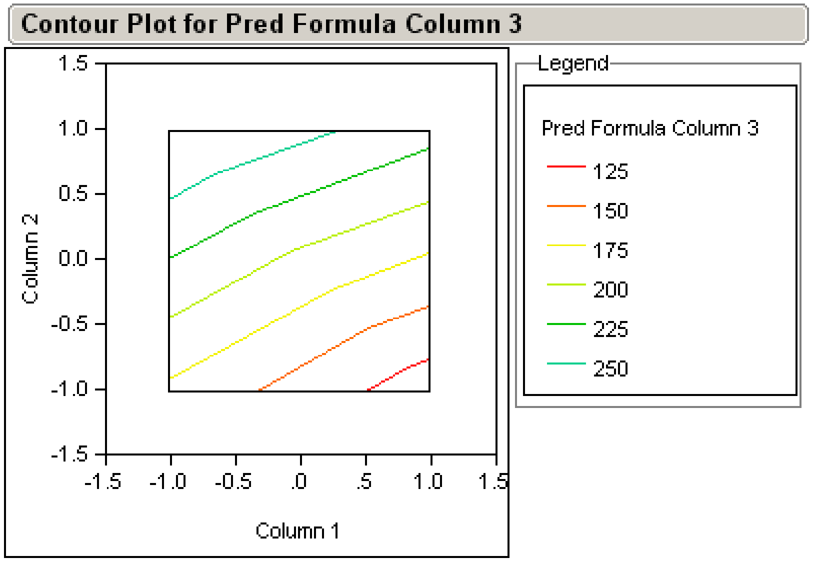
Regression Analysis for the Main Effect influence

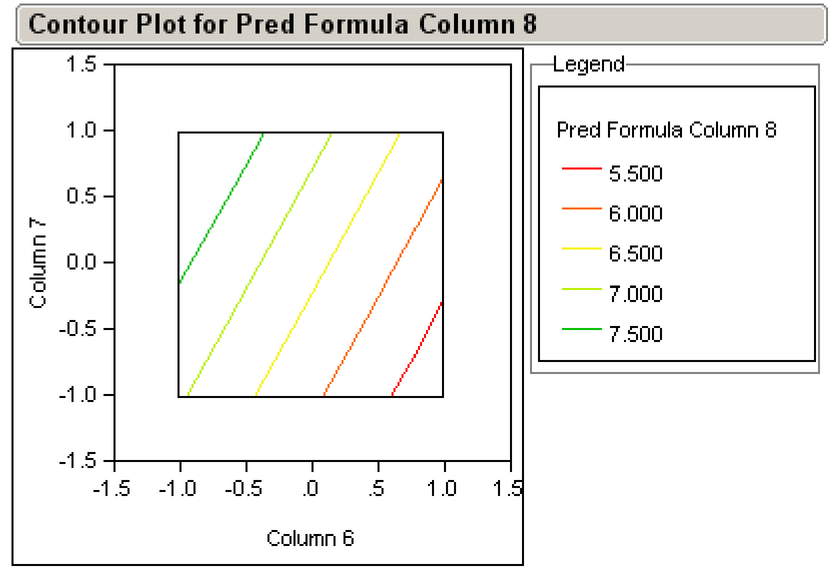
	Coef	Std Err	t Stat	P-val	Lower 95	Upper 95
Intercept	197.13	7.88	25.01	0.00	181.00	213.25
X Var 1	-27.50	7.88	-3.49	0.00	-43.62	-11.38
X Var 2	56.88	7.88	7.21	0.00	40.75	73.00

Regression Analysis for the Variance Influence

	Coef	Std Err	t Stat	P-val	Lower 95	Upper 95
Intercept	6.77	0.78	8.70	0.00	4.77	8.77
X Var 1	-0.82	0.78	-1.05	0.34	-2.82	1.18
X Var 2	0.69	0.78	0.88	0.42	-1.31	2.69

Graphical *Combined* DOE Approach





Optimal Solution:

Overlaying both plots (for location and variation) we seek to Minimize simultaneously Yield and Variation.

Jointly applying the two above (cols. 3 & 8).

The Optimum is around (1, -1), yielding

Estimated Minimum Output = 113; Min Variation = 5.3

Estimated Yield:

Y = 197.12 - 27.5X1 + 56.9X2

Y (1, -1) = 112.72

Estimated Variation:

Y = 6.77 - 0.82X1 + 0.69X2

Y (1, -1) = 5.26

Pan-American Advanced Studies Institute

Synopsis of Program:

The Pan-American Advanced Studies Institutes (PASI) Program is a jointly supported initiative between the Department of Energy (DOE) and the National Science Foundation (NSF). Pan-American Advanced Studies Institutes are short courses ranging in length from ten to twenty-one days, involving lectures, demonstrations, research seminars, and discussions at the advanced graduate, post-doctoral, and junior faculty level.

PASIs aim to disseminate advanced scientific and engineering knowledge and stimulate training and cooperation among researchers of the Americas in the mathematical, physical, and biological sciences, the geosciences, the computer and information sciences, and the engineering fields. Proposals in other areas funded by NSF may be considered on an ad hoc basis as long as they are multidisciplinary; in this case, lead investigators must consult with the PASI Program before proposal submission.

Pan-American Advanced Studies Institute

- US-Latin American Scientists/Researchers
- Modeling of Environmental Problems
- Modelers: statistics & applied math (O.R.)
- Environmental Science Specialists
- From USA: EPA, GLRC, Other Universities
- From LA: Mexico, Brazil, Argentina, Chile, Colombia, Ecuador, Puerto Rico, others
- Via the Juarez Lincoln Marti Int'l Ed. Project
  - http://web.cortland.edu/matresearch

Conclusions

- DOEs are complex methodologies
  - Size and interactions are serious issues
- Existing methods, not fully compliant
  - But promise, if worked around
- Some Models are useful
  - For strategic and tactical decisions
  - In crisis, and to assess/avoid them
  - In theoretical and applied studies
- A PASI for Latin America in preparation