Virtual Experiments & Generating Hypotheses Using Computer Simulations

Kathleen M. Carley
Carnegie Mellon University
kathleen.carley@cs.cmu.edu

Center for Computational Analysis of Social and Organizational Systems

http://www.casos.cs.cmu.edu/

What Is a Virtual Experiment?

- An experiment run using a computer simulation model
- Standard experimental design procedures
  - Identify variables (parameters)
  - Define a set of experimental cells
  - Run a series of virtual experiments - Rerun simulation in each cell multiple times if stochastic elements
  - Statistically analyze results - Locate the best fit model (typically nonlinear).
- Typically done for models with large numbers of parameters.
- Basic Goal: Map inputs to outputs in area of concern
- Alternative Name: Local Response Surface Analysis

June 13, 2006 2

Carnegie Mellon

Types of Virtual Experiments

- Monte Carlo.
  - Parameters not directly varied are 'factored out' using a Monte Carlo approach (large number of runs across which all other parameters are randomly selected from acceptable options). Result is point estimate of behavior for each cell in the experimental design.
- Empirical Monte Carlo.
  - Model is initialized with 'real' data. Then each cell in experimental design is estimated using a Monte-Carlo approach over values not set by 'real' data.
- Pattern matching.
  - A set of simulations are run that correspond exactly to each 'real' case.
- Proof of concept.
  - Model is run once each under a set of conditions to demonstrate what can happen.

Carnegie Mellon

Key Questions Virtual Experiments Answer

- **Likelihood**
  - **Is A likely?**
  - ***Will mimicry tend to improve performance?***
- **Relative impact**
  - **Does A or B have a greater impact?**
  - ***Does task or structure have a greater impact on performance?***
- **Possibility**
  - **Is A possible?**
  - ***Can organizations improve performance if individuals learn and the organization structurally adapts?***
- **Sufficiency & necessity**
  - **Is A sufficient or necessary?**
  - ***Is forgetting necessary to decrease interaction?***

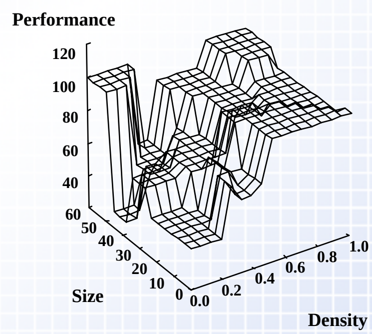
Describing a Virtual Experiment

- Create a table
  - With 3 columns – variable, values, the number of values
  - With one row per variable
  - Last row contains
    - The number of repetitions per cell
    - The time periods each repetition is run for
    - The total number of cells (this is a nxmxz = p design)

Variable	Values	Number
Age ranges	0-20, 20-40, 40-60, 60+	4
Gender	Male, female	2
New Ideas	1,2,10	3
Repetitions per cell = 30	Time periods = 100	Design = 4x2x3=24 cells

Virtual Experiments generate response surfaces

- Response surfaces are sometimes referred to as landscapes
- Illustrative Landscape for OrgAhead

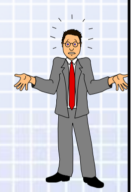


Caveat: For natural organizations 3 dimensions are not enough

Carnegie Mellon

RESPONSE SURFACE METHODOLOGY

- One Definition:
  - A statistical method
  - That uses quantitative data from appropriate Virtual Experiments
  - To determine and simultaneously solve multivariate equations
- Another Definition:
  - Statistical mapping of a Simulation Model's Outputs to Inputs of Concern within a region
- Response surface analysis provides a means for optimization of process model
- Selection of parameters and values for virtual experiment is critical
- Each virtual experiment design has its own set of advantages and disadvantages



CASOS

June 13, 2006 7

Carnegie Mellon

Response Surface Modeling

- Mapping out the performance surface of a computational model.
- Typically done for models with large numbers of parameters.
- Typically done using Monte-Carlo estimation of surface.
- Select sets of parameters that span the surface are of concern.
  - Cover extreme points.
  - Known inflection points.
- Run series of virtual experiments.
- Locate the best fit model (typically non-linear).

CASOS

June 13, 2006 8

Carnegie Mellon

Procedure for Generating Hypotheses

- 1. Identify core variables / parameters – limit the number if possible
- 2. Explore the parameter space.
- 3. Set non-core variables – limit the number is possible
- 4. Define parameter levels
  - If too broad, optimization cannot be defined
  - Levels may be restricted by cost, physical limits or regulation, so that optimum may be outside levels tested
- 5. Select test samples
  - Once parameters are set, first run a test at mid point to test levels
- 6. Select experimental design method
- 7. Choose Data analysis method
- 8. Run simulations in virtual experiment.
  - Select test samples
  - Select experimental design method
- 9. Statistically analyze results.
- 10. Formulate hypotheses.

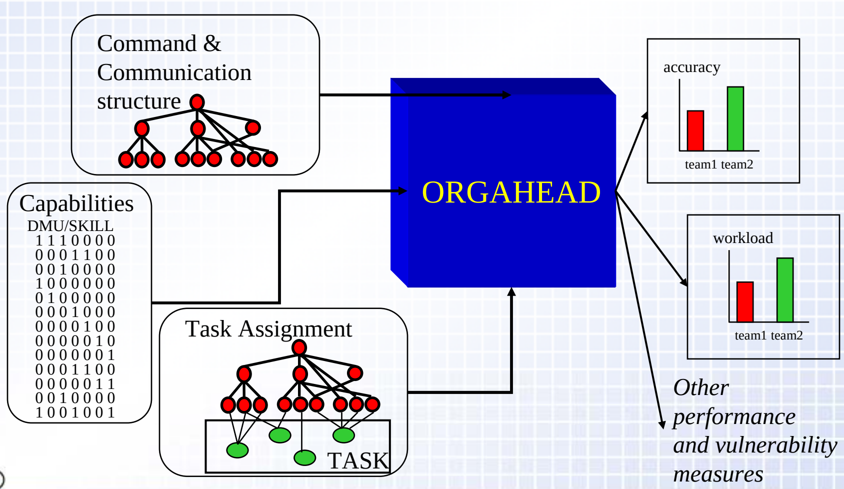
Assumes you have a working simulation model

CASOS

June 13, 2006 9

Carnegie Mellon

ORGAHEAD Overview



CASOS

June 13, 2006 10

Carnegie Mellon

1. Identify Core Variables

- Core variables are the parameters of concern
- Parameters or model modules which are expected to be highly relevant in affecting the dependent variable
- Not all model parameters are core

Example:

- ORGAHEAD --- task complexity

CASOS

Carnegie Mellon

Which Variables/Parameters to Consider - Assumptions

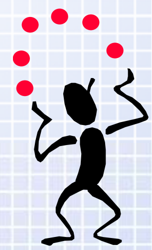
- **Critical parameters are known**
- **Region of interest, where parameter levels are known or considered critical**
- **Parameters vary continuously through-out the experimental range tested**
- **A mathematical function relates the parameters to the response (outcome)**
- **The response defined by this function is a smooth curve**

CASOS

Carnegie Mellon

Choosing Parameters

- Easy to generate too much data to analyze
- Most data generable by simulations is never analyzed
- Exhaustive analysis of response surface not possible
- Balance – the number of things varied, the number of replications, and the number of outcomes being observed
- Current state of the art
  - Run more small VE rather than one large one
  - New directions
    - Data farming environments
    - Ants

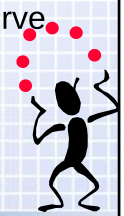


CASOS

Carnegie Mellon

Which Parameters to Consider - Assumptions

- Critical parameters are known
- Region of interest, where parameter levels are known or considered critical
- Parameters vary continuously through-out the experimental range tested
- A mathematical function relates the parameters to the response (outcome)
- The response defined by this function is a smooth curve



CASOS

Carnegie Mellon

2. Explore the Parameter Space

- Define which values for each parameter will be explored
- The choice should reflect
  - Concerns
  - Expectations about impact on system level behavior
  - Known points of relevance
- Choose 2 or more values for each parameter
  - Preferably 3

Example

- Task complexity
  - Low, medium and high complexity

CASOS

Carnegie Mellon

3. Set Non-core Variables

- Non core variables should be
  - Randomly set
  - Fixed at a level needed for the analysis
  - Set to match conditions known to be true of human groups

Example

- Fixed parameter: size - set randomly, but doesn't change

CASOS

Carnegie Mellon

Considerations for Steps 2 and 3

- You are creating a virtual experiment
- Define parameter levels
  - If too broad, optimization can not be defined
  - Levels may be restricted by cost, physical limits or regulation so that the optimum is outside of levels tested
  - Once parameter levels are set, run a test at midpoint on each parameters value to test the levels

CASOS

June 13, 2006 17

Carnegie Mellon

Choosing the Design for the Virtual Experiment

- **Response predictions always have some degree of uncertainty**
- **Design should help ensure that the predictions are reasonable throughout the experimental range**
- **Try to pick points to ensure that a uniform prediction error is obtained by using an experimental design that fills the area of interest**
- **The final choice of the experimental design is affected by the shape of the area of interest**

CASOS

June 13, 2006 18

Carnegie Mellon

Experimental Design

- Select an experimental design
- Commonly recommended:
  - Central Composite designs
  - Box Behnken design
  - Draper and Lin, minimal central composite designs
  - Morris designs
    - See engineering statistics handbook e.g. ch. 5.3 -
    - http://www.itl.nist.gov/div898/handbook/pri/section3/pri3.htm
- Suitable for fitting a second order process model
    - Order:
    - 1) Completely linear
    - 2) Square or two-way interactions + 1
    - 3) Cubed or three-way interactions + 2) + 1

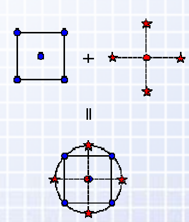
CASOS

Carnegie Mellon

Central Composite

- Also known as Box-Wilson Central Composite Design
- Contains imbedded factorial or fractional factorial design with center points
- Augmented with a group of `star points'
- This allows estimation of curvature
- Note – if the distance from the center of the design space to a factorial point is ±1 unit for each factor, the distance from the center of the design space to a star point is ± α with | α | > 1
- The value of α depends on the properties desired for design and on the number of factors involved
- The number of center point runs the design contains also depends on properties required for the design

Central Composite Design for Two Factors

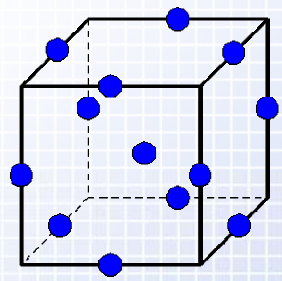


CASOS

Box Behnken

- An independent quadratic design
- Does not contain an embedded factorial or fractional factorial design
- Treatment combinations are at the midpoints of edges of the process space and at the center
- Designs are rotatable (or near rotatable) and require 3 levels of each factor
- Designs have limited capability for orthogonal blocking compared to central composite designs

Design for 3 Factors



Extreme Point Analysis

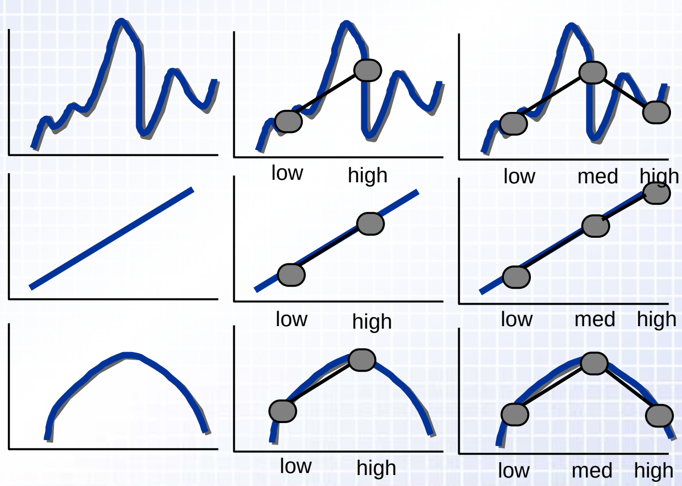
- Examine the extreme points
  - E.g., groups of 2
  - Learning is off
- Rational
  - Such points might have a mathematical equivalent
  - Trace may be easier to follow
  - May be useful for face validity

LMH Rule

You can't look for non-linearities with only two points

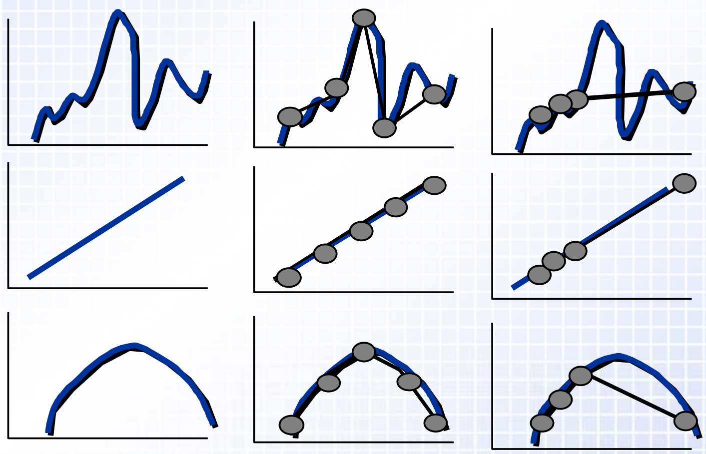
- When setting up a virtual experiment for non-categorical parameters use more than two points
- Examples
  - size - 10, 20, 30, 40
  - density - .1 , .15, .2, .25
- Span the space
- Pick values known to correspond to “physical world” data
- Pick values distant enough for results to be distinguished
- Run a small trial run first
- The number of points depends on statistical significance considerations, computation time, space constraints, and theory

Examples



Carnegie Mellon

Equal or Unequal Spacing



CASOS

Carnegie Mellon

ORGAHEAD Parameter Space

Choosing parameters and values defines virtual experiment

Parameter	Categories
Task limit	20,000 and 80,000
Task complexity	binary and trinary
Task information	7 and 9
Agent ability	5 and 7
Stressors	Stable and periodic
Unit Size	9, 12, 18, and 36
Shake-ups	1, 2, 3 and 4

Table 1: Summary of Parameters

512 cells

CASOS

Carnegie Mellon

Virtual Experiment Design Considerations

- Generally, the area of interest is determined by the ranges of the independent variables (parameters)
  - E.g., if the region is cubical (in coded values of X (the parameters)) then the best design is face centered.
- If you need the precision of the predictions to be independent of the direction from the center – then the region is "spherical" and you should use a Box-Behnken design
- Box-Behnken designs exclude the corners where all variables are simultaneously at the extremes
- This permits a wider range of individual variables
- If the shape of the virtual experiment is neither spherical or cubical and has strong constraints then the region may be an irregular tetrahedron and will require a special design

CASOS

June 13, 2006 27

Carnegie Mellon

Review the Design for Operability

- Review runs for operability
  - How much space and time will this take
- Note – runs that set the variables at maximum or minimum values may not work
  - Programming errors
  - Time and space issues
- Randomization can be "adjusted" to schedule these runs early to allow for adjustments
- Do exploratory testing of potentially troublesome runs before running the entire virtual experiment on auto-pilot

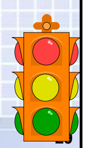
CASOS

June 13, 2006 28

Carnegie Mellon

How to Avoid Blunders

- Try out a simple case first
- Execute experiment with care
- Small statistical designs are susceptible to errors as every run estimates more than on effect
- Record results for all runs
- Plan for analysis from the start
- Statistical analysis (regression) is the basis for most analytic procedures
- Make sure the results “make sense”



CASOS

June 13, 2006

Carnegie Mellon

Limitations to Response Surface Modeling

- POLYNOMIAL MODELS DO A POOR JOB OF PREDICTING RESPONSE OUTSIDE THE REGION OF EXPERIMENTATION
- LARGE VARIATIONS IN THE FACTORS CAN BE MISLEADING (ERROR, BIAS, NO REPLICATION)
- CRITICAL FACTORS MAY NOT BE CORRECTLY DEFINED OR SPECIFIED
- RANGE OF LEVELS OF FACTORS TO NARROW OR TO WIDE - OPTIMUM CAN NOT BE DEFINED
- LACK OF USE OF GOOD STATISTICAL PRINCIPLES
- OVER-RELIANCE ON COMPUTER - MAKE SURE THE RESULTS MAKE GOOD SENSE
- FEW TECHNIQUES FOR TIME VARIANT SURFACES
- DIFFICULT FOR HIGHLY RUGGED SURFACE
- CANNOT BE DONE EXHAUSTIVELY FOR LARGE COMPLEX MODELS

CASOS

June 13, 2006 30

Carnegie Mellon

4. Run Simulations in Virtual Experiment

- Run simulations
- If stochastic - run multiple simulations per cell - replications
  - Example - 40 replications per cell
- If deterministic - run one simulation per cell
- Comment: virtual experiments generate LOTS of data
  - Example
    - Virtual experiment described resulted in 20480 data observations at each point in time

CASOS

June 13, 2006 31

Carnegie Mellon

5. Statistically Analyze Results

- Lots of data - multiple explorations

Example

- Exploration 1: the impact of meta-adaptation strategies on performance
- Exploration 2: the impact of meta-adaptation strategies on the C3I structure

CASOS

June 13, 2006 32

Carnegie Mellon

General Approach to Response Surface

- Approach
  - First run linear regression with all parameters (level 1)
    - Don't expect high R2 as the point is a non-linear system
  - Now add in all simple products (level 2)
    - R2 should be improved
  - Now add in all cross terms with 3 variables (level 3)
    - R2 should be improved
  - …
- As increase number of terms have at least 25 cases per term
- Difficulties
  - With many multi-agent systems you run out of storage space before you get to level 3
  - It is easy to over-run existing stat packages
  - Even if characterize the system at time x a different response may exist at time y

CASOS

June 13, 2006 33

Carnegie Mellon

What statistical method ?

- **If all variables are continuous you can use regression analysis**
- **Search for equation with minimum sum of squared errors**
- **If you want to "punish" outliers more use minimum sum of squared errors**
- **If you want to "punish" outliers less use the sum of the absolute value of the errors**

CASOS

June 13, 2006 34

Carnegie Mellon

Set of Equations

- Assume 1 dependent (Y) and 2 independent X and Z
- 1st order (linear) (all possible combinations of linear terms)
  - Y=a + b1Y + b2Z
  - Y=a + b1Y
  - Y=a + b2Z
- 2nd order (all possible combinations of linear and 2nd order terms)
- 3rd order (all possible combinations of linear and 2nd order and 3rd order terms)
- Note – we typically don't need to go over 3rd order
- At this point we don't worry about fancier forms with logs or exponentials

- Y=a + b1Y + b2Z + b3YZ + b4Y2 + b5Z2
- Y=a + b1Y + b3YZ + b4Y2 + b5Z2
- Y=a + b2Z + b3YZ + b4Y2 + b5Z2
- Y=a + b1Y + b2Z + b3YZ + b4Y2
- Y=a + b1Y + b3YZ + b4Y2
- Y=a + b2Z + b3YZ + b4Y2
- Y=a + b1Y + b2Z + b3YZ + b5Z2
- Y=a + b1Y + b3YZ + b5Z2
- Y=a + b2Z + b3YZ + b5Z2
- Y=a + b1Y + b2Z + b4Y2 + b5Z2
- Y=a + b1Y + b4Y2 + b5Z2
- Y=a + b2Z + b4Y2 + b5Z2
- Y=a + b1Y + b2Z + b3YZ
- Y=a + b1Y + b3YZ
- Y=a + b2Z + b3YZ
- Y=a + b1Y + b2Z + b5Z2
- Y=a + b1Y + b5Z2
- Y=a + b2Z + b5Z2
- Y=a + b1Y + b2Z + b4Y2
- Y=a + b1Y + b4Y2
- Y=a + b2Z + b4Y2
- Y =a + b3YZ + b4Y2 + b5Z2
- Y = a +b3YZ + b4Y2
- Y = a +b3YZ + b5Z2
- Y =a + b4Y2 + b5Z2
- Y =a + b3YZ
- Y =a + b5Z2
- Y =a + b4Y2

CASOS

Carnegie Mellon

Example Statistical Analysis

Predictor	Coefficient	p value
intercept	0.000000	1.000
Task limit	0.031853	0.000
Task complexity	-0.024068	0.000
Environmental stressors	-0.014568	0.027
Unit size	0.170226	0.000
Agent ability	0.265205	0.000
Task information	0.091118	0.000
Shake-ups	-0.012299	0.063

R2 (adj) = 10.9%, df = 7, 20472, p<0.001

Table 2. Standardized Regression for Performance.

Resources
Size
Task needs

CASOS

Carnegie Mellon

6. Generate Hypotheses

- Turn strong statistical results into hypotheses
- Sensitivity analysis set limits on these hypotheses
- Any types of hypotheses possible:
  - Simple predictions of correlations
  - Statements about the shape of a distribution
  - Statements of relative impact
  - Quantitative statements about impact

Carnegie Mellon

Illustrative Hypotheses

- The more limited the task the higher the performance
- The higher the level of task complexity the lower the level of performance
- As unit size increases so does unit performance
- Hiring capable personnel should improve performance more than increasing the amount of available information
- Increasing personnel capability will have a 1.5 time greater performance improvement than will increasing organization size
  - Corollary: organizations who retrain, rather than hire new personnel should do better

Carnegie Mellon

Theory Development

Generation of hypotheses

\+ Model development

Theory development

CASOS

Carnegie Mellon

Summary

- Computational models as hypotheses generators
- 6 key steps in generating hypotheses
  - Virtual experiment - statistical analysis - hypotheses
- Basis of theory development
- Part of procedure for using scientific method for team design

CASOS