AlaSim International

6-7, May 2014 • Huntsville, Alabama • USA

Organized By AMSC Alabama Modeling and Simulation Council

Model Verification and Validation Methods

Mikel D. Petty, Ph.D.

University of Alabama in Huntsville

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

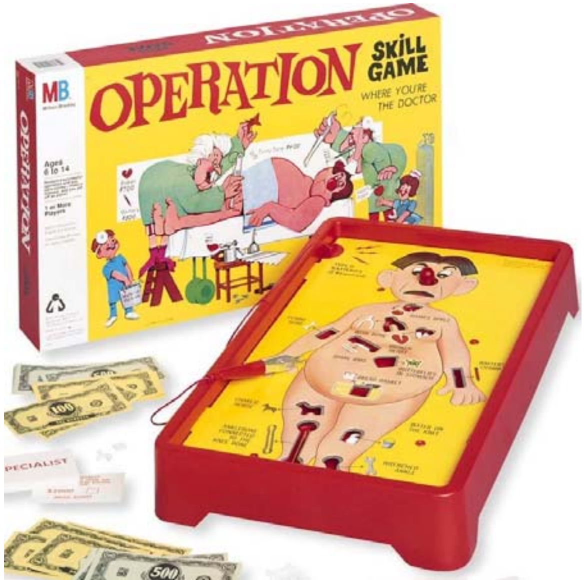
Outline

- Motivation and introduction
- Definitions and concepts
- A survey of verification and validation methods
  - Informal methods
  - Static methods
  - Dynamic methods
  - Formal methods
- Case studies
  - Validation using confidence intervals
  - Validation using a statistical hypothesis test
  - Comparing real and simulated missile impact data
- Summary

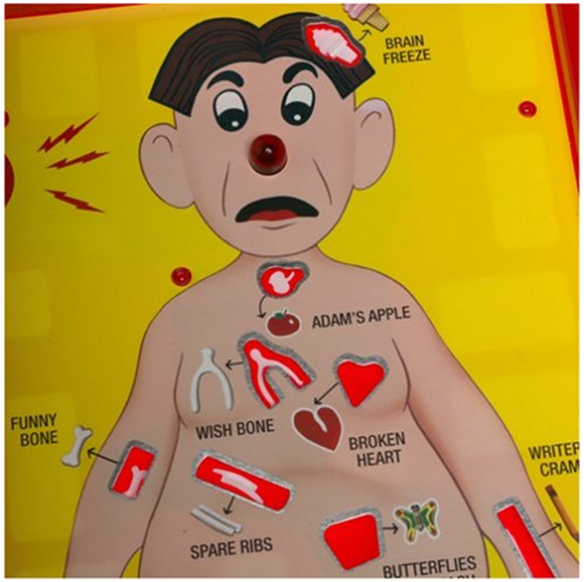
Motivation and introduction

Motivating example: *Operation*

- Milton Bradley *Operation* game
  - Remove plastic “ailments” from “Cavity Sam”
  - Avoid touching tweezers to perimeter of opening
- Suitable for training surgeons?



Operation game equipment



Representation of human anatomy?

Motivating example: *Zero-flight time simulators*

- Zero-flight time simulators
  - Simulator recreates aircraft controls, flight dynamics
  - Airline pilots train on new aircraft type in simulator
- Suitable for training pilots?

Flight simulator cockpit

Participants in pilot’s first flight in the aircraft type?

Motivation and learning objectives

- Motivation
  - VV&A essential to credible and reliable use of M&S
  - Full range of V&V methods not widely known
  - V&V execution depends on context and application
- Learning objectives
  - Define and compare verification and validation
  - Define and contrast categories of V&V methods
  - List V&V methods within each category
  - For select V&V methods, explain each method and state what types of models it applies to
  - State important findings from V&V case studies

There's more to V&V than "that looks about right".

Definitions and concepts

Concepts

- Model: representation of something else
- Simulation: executing a model over time

$$R = 2.59 \times \sqrt[4]{\sigma \times \left(\frac{\log^{-1}\left(\frac{ERP_t}{10}\right) \log^{-1}\left(\frac{G_r}{10}\right) \log^{-1}\left(\frac{MDS_r}{10}\right)}{\log^{-1}\left(\frac{FEL_r}{10}\right) F_t^2} \right)}$$

Model

Simulation

Both

Definition

Model. A physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process. [DOD, 1996] [DOD, 2009]

- Representation of something else, often a “real-world” system
- Some aspects of the modeled system are represented in the model, others not

Definition

Simulation. Executing a model over time. Also, a technique for testing, analysis, or training in which real world systems are used, or where a model reproduces real world and conceptual systems. [DOD, 1996] [DOD, 2009]

Alternative uses of term (to be avoided)

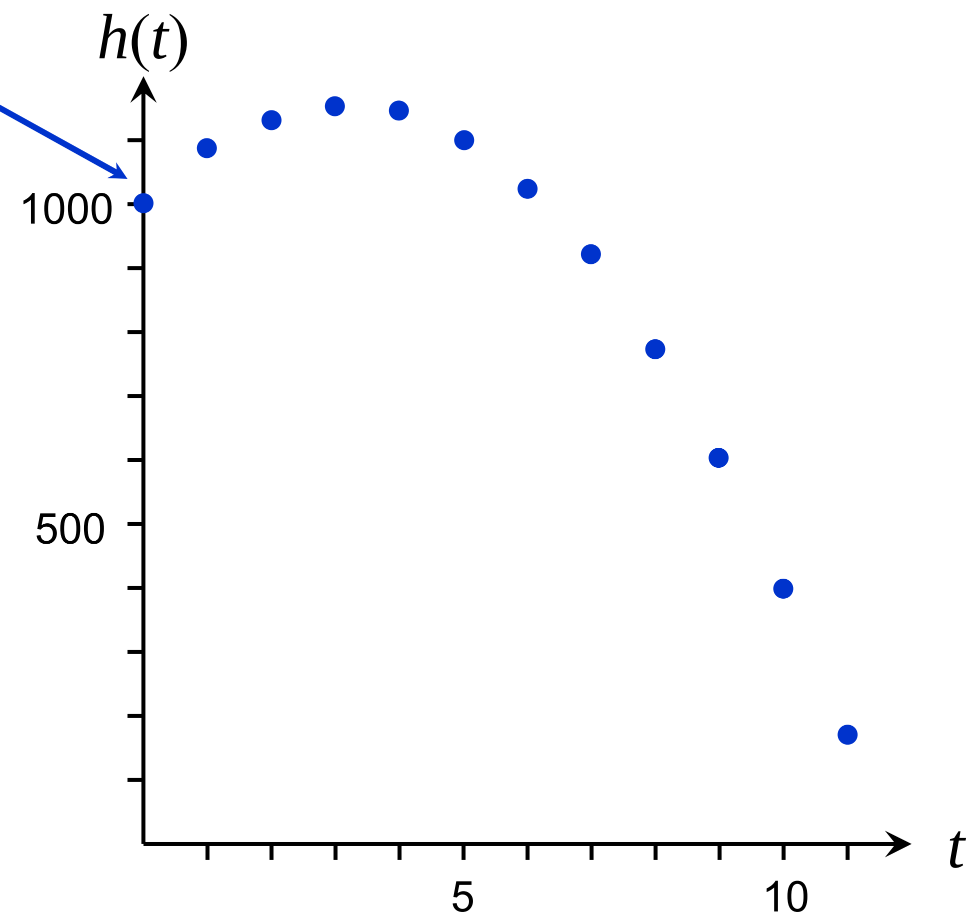
- A large composite model
- Software implementation of a model

Example: Height under gravity

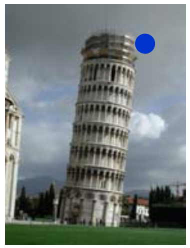
Model: $h(t) = -16t^2 + vt + s$ Data: $v = 100$, $s = 1000$

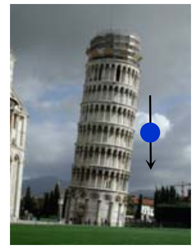
t	$h(t)$
0	1000
1	1084
2	1136
3	1156
4	1144
5	1100
6	1024
7	916
8	776
9	604
10	400
11	164

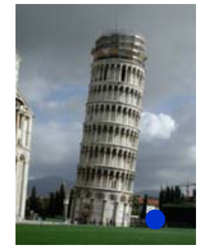
Start state



Simulation vs reality







Real-world system in start state → **Time** / *Physics* → Real-world system in end state

Modeling

Initialization

Interpretation

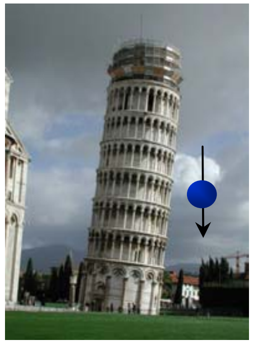
Validation

Model in start state → ***Simulation*** / *Computation* → Model in end state

$$h(t) = -16t^2 + vt + s$$
$$1000 = -16(0)^2 + 100(0) + 1000$$

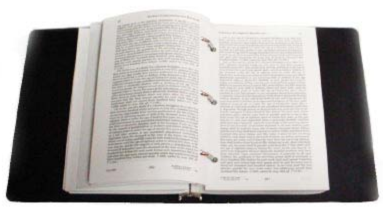
$$h(t) = -16t^2 + vt + s$$
$$0 = -16(11.63)^2 + 100(11.63) + 1000$$

Background definitions, 1 of 2



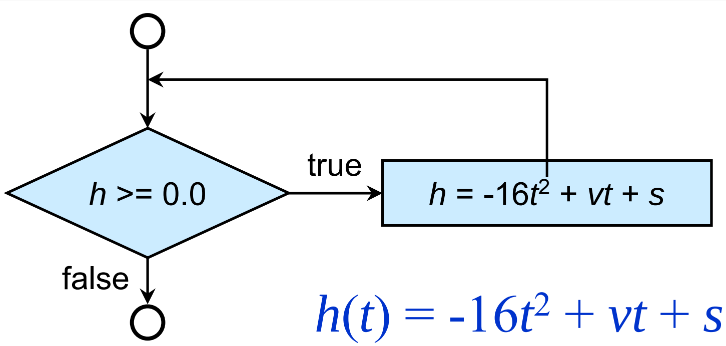
Simuland

- Real-world system
- Thing to be simulated



Requirements

- Intended uses
- Needed validity, resolution, scale



$$h(t) = -16t^2 + vt + s$$

Conceptual model [Banks, 2010]

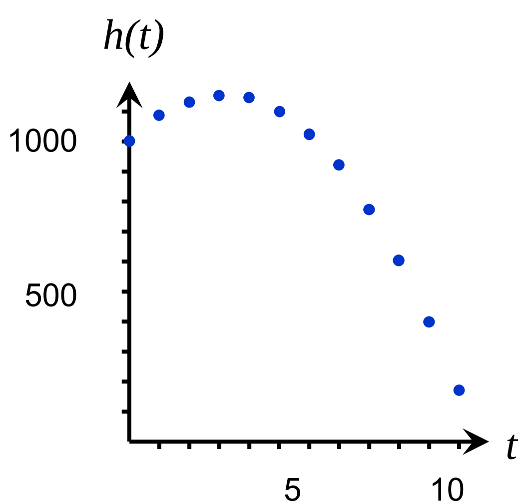
- Simuland components, structure
- Aspects of simuland to model
- Implementation specifications
- Use cases
- Assumptions
- Initial model parameter values

Background definitions, 2 of 2

```
/* Height of an object moving in gravity.  */
/* Initial height v and velocity s constants. */
main()
{
  float h, v = 100.0, s = 1000.0;
  int t;
  for (t = 0, h = s; h >= 0.0; t++)
  {
    h = (-16.0 * t * t) + (v * t) + s;
    printf("Height at time %d = %f\n", t, h);
  }
}
```

Executable model

- Computer software
- Implemented conceptual model



t	*h(t)*
0	1000
1	1084
2	1136
3	1156
4	1144
5	1100
6	1024
7	916
8	776
9	604
10	400
11	164

Results

- Output of model
- Produced during simulation

Definition

Verification. The process of determining that a model implementation and its associated data accurately represents the developer's conceptual description and specifications. [DOD, 2009]

- Transformational accuracy
  - Transform specifications to code
- Software engineering quality
  - Software engineering methods apply
- Summary question
  - Is the model coded right? [Balci, 1998]

Definition

Validation. The process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model. [DOD, 2009]

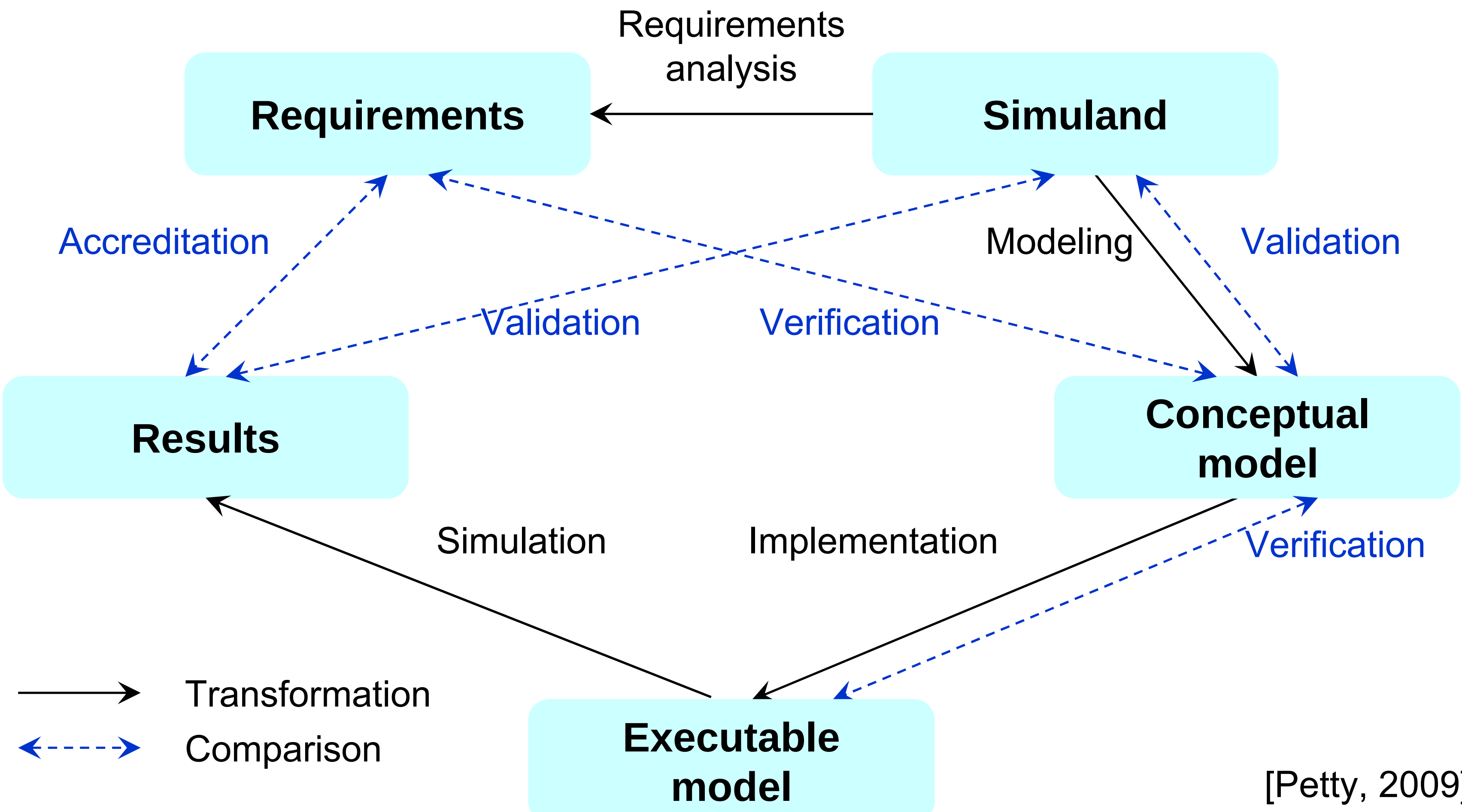
- Representational accuracy
  - Recreate simuland with results
- Modeling quality
  - Special validation methods needed
- Summary question
  - Is the right model coded? [Balci, 1998]

Definition

Accreditation. Official certification [by a responsible authority] that a model or simulation is acceptable for use for a specific purpose. [DOD, 2009]

- Official usability for specific purpose or function
  - Management decision, not technical process
  - Not a blanket or general-purpose approval
- Accrediting (or accreditation) authority
  - Agency or person responsible for use of model
  - Normally not model developer
- Summary question
  - Is the model the right one for the job? [Petty, 2010]

VV&A comparisons



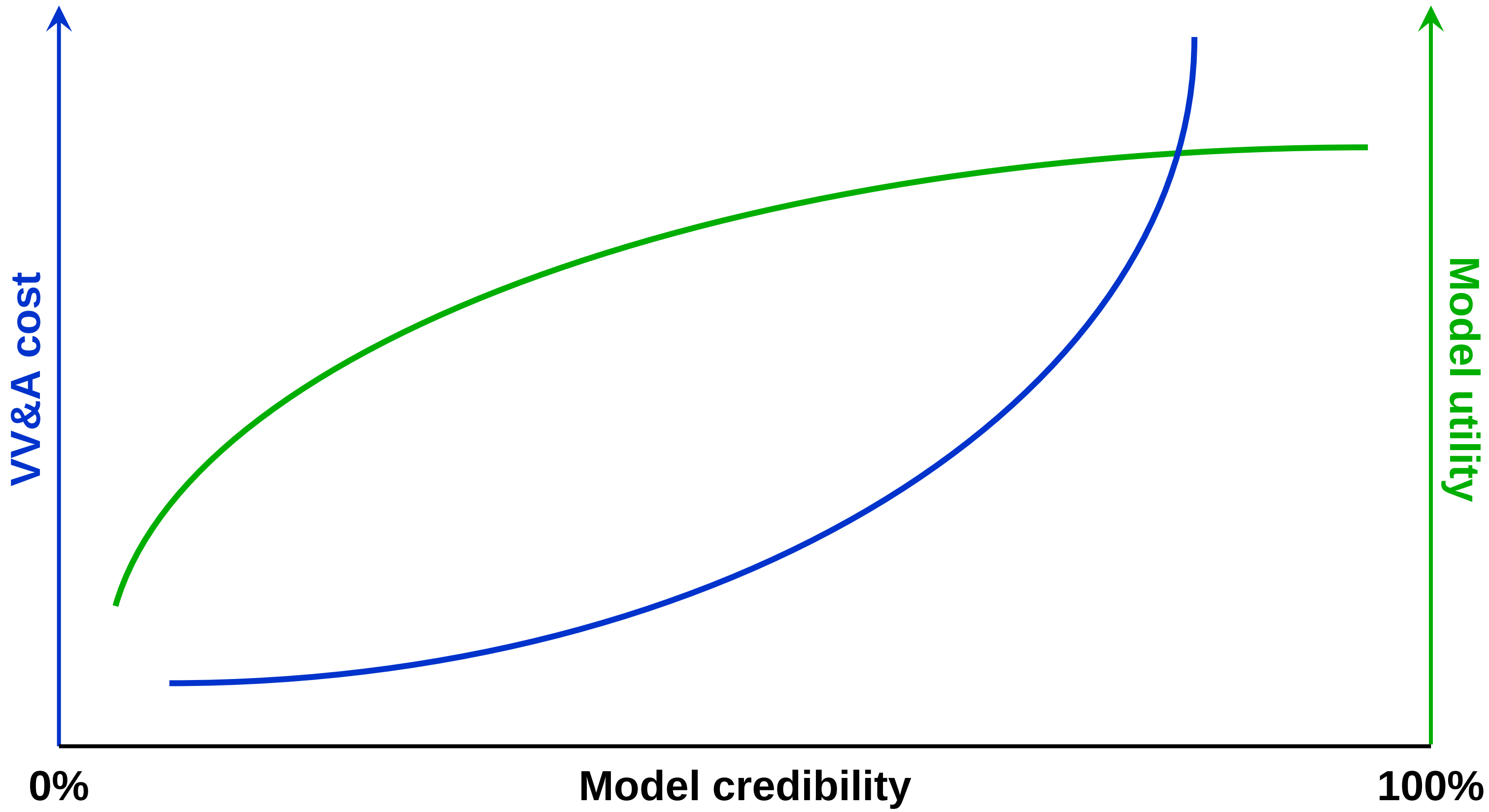
[Petty, 2009]
[Petty, 2010]

VV&A errors and risks [Balci, 1981] [Balci, 1985] [Balci, 1998]

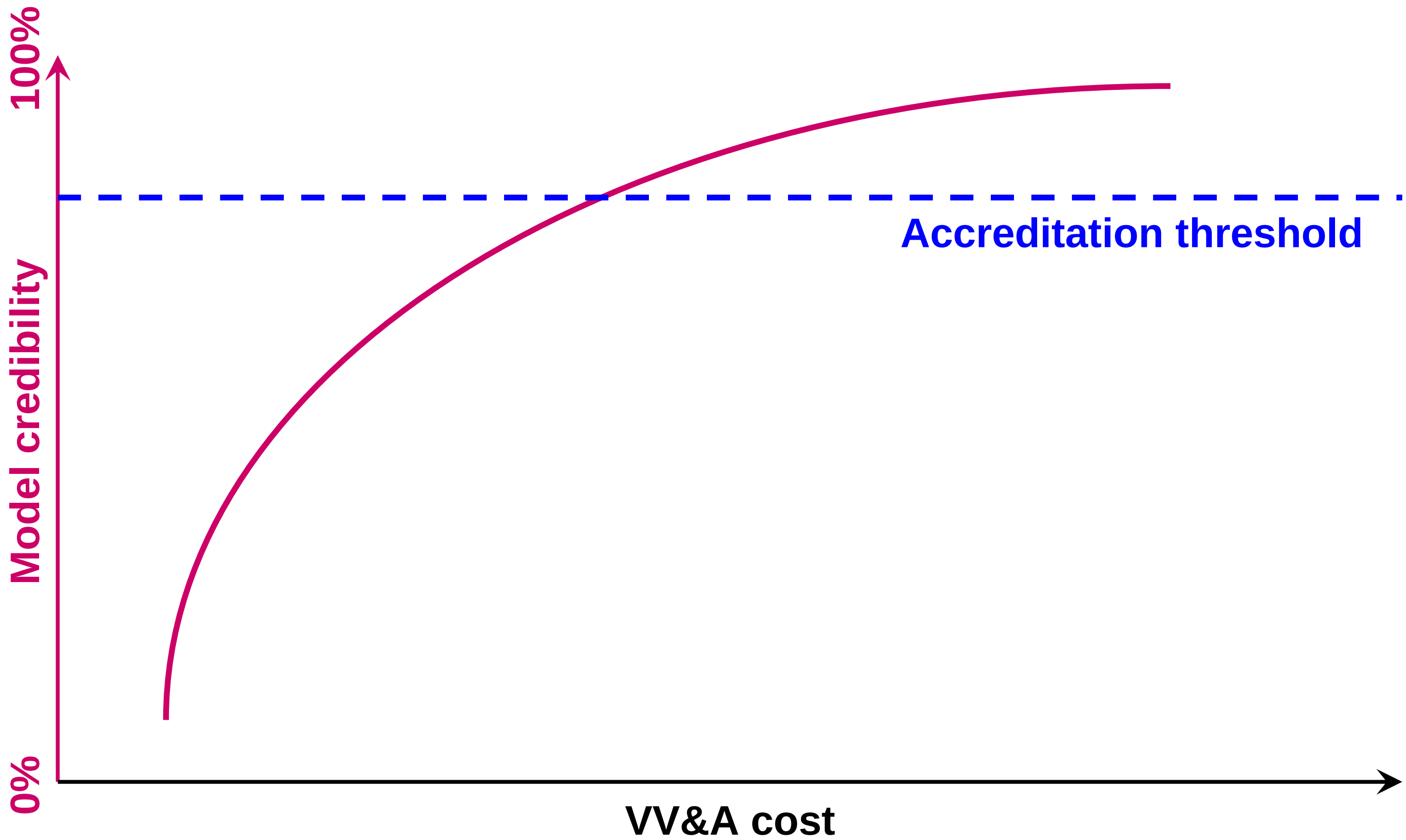
	Model valid	Model not valid	Model not relevant
Model used	Correct	**Type II error** Use of invalid model; Incorrect V&V; Model user's risk; **More** serious error	**Type III error** Use of irrelevant model; Accreditation mistake; Accreditor's risk; **More** serious error
Model not used	**Type I error** Non-use of valid model; Insufficient V&V; Model builder's risk; **Less** serious error	Correct	Correct

Credibility, cost, and utility

[Shannon, 1975] [Sargent, 1996] [Balci, 1998] [Sargent, 2000]



How much VV&A is enough?



A survey of verification and validation methods

V&V methods

- Many available, ~85 in 1998 [Balci, 1998], more since
- Different purposes, advantages

Informal	Static	Dynamic	Formal
-Audit -Desk checking -Documentation Checking -Face validation -Inspections -Reviews -Turing test -Walkthroughs	-Cause-Effect Graphing -Control Analysis -Data Analysis -Fault/Failure Analysis -Interface Analysis -Semantic Analysis -Structural Analysis -Symbolic Evaluation -Syntax Analysis ...	-Acceptance Testing -Alpha Testing -Assertion Checking -Beta Testing -Bottom-up Testing -Comparison Testing -Statistical Techniques -Structural Testing -Submodel/Module Testing ...	-Induction -Inductive Assertions -Inference -Logical Deduction -Lambda Calculus -Predicate Calculus -Predicate Transformation -Proof of Correctness

[Balci, 1998]

V&V methods

- > 100 V&V methods
- Organized into categories [Balci, 1998]
  - Informal
  - Static
  - Dynamic
  - Formal
- Similarities
  - Forms of testing
  - Involve comparisons [Petty, 2009] [Petty, 2010]
- Differences
  - What is being compared
  - Degree of formality and quantitativeness
  - Appropriate applications

Factors affecting the selection of V&V methods

$$h(t) = -16t^2 + vt + s$$

Model

Data

```
01010100010100101
10010101000100101
10101011101010111
10110010100101000
10010101101010101
```

Certain methods apply best to specific types of model

Certain methods require specific types or amounts of data

Method

Some methods require more resources than others

Some methods require specific skills to apply

Resources

Skills

$$t = \frac{\left(\overline{X}(n) - \mu\right)}{\sqrt{\frac{S^2(n)}{n}}}$$

Problems can arise if the factors conflict

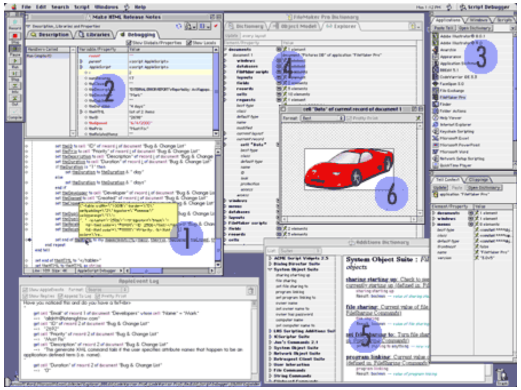
Informal methods

Informal V&V methods

- Characteristics
  - Methods that rely heavily on Subject Matter Expert (SME) expertise and evaluation
  - More often qualitative and subjective
  - More often performed by SMEs
- Example informal V&V methods
  - Inspection
  - Face validation
  - Turing test

Inspection (verification)

- Organized teams of developers, testers, and users inspect artifacts
- Compare
  - Requirements to conceptual model
  - Conceptual model to executable model
- Errors found by manual examination
- General software verification method

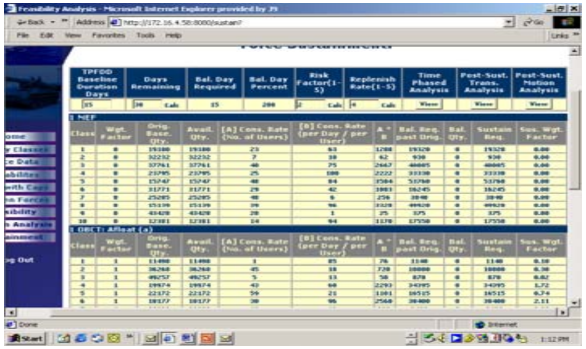


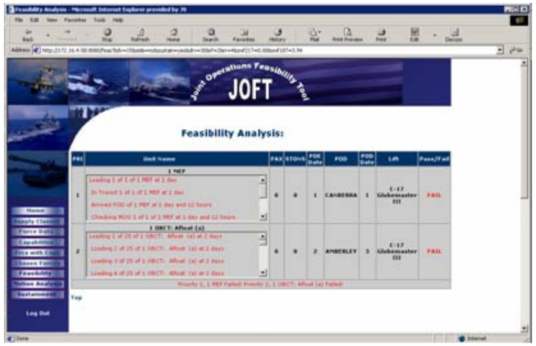
Face validation (validation)

- SMEs, modelers, and users observe model execution and/or examine results
- Compare results to simuland behavior, as understood by SMEs
- Assessment
  - Model validity evaluated subjectively
  - Based on expertise, estimates, and intuition
- Comments
  - Frequently used because of simplicity
  - Often used when user interaction important
  - Clearly better than no validation

Face validation example

- Joint Operations Feasibility Tool [Belfore, 2004]
  - Assess deployment transportation feasibility
  - Assess logistical sustainment feasibility
- Validation process
  - Special scenarios exercise full range of capabilities
  - 20 SMEs with extensive experience evaluated model
  - Assessments elicited via written questionnaires
- Process structure addressed face validation limits

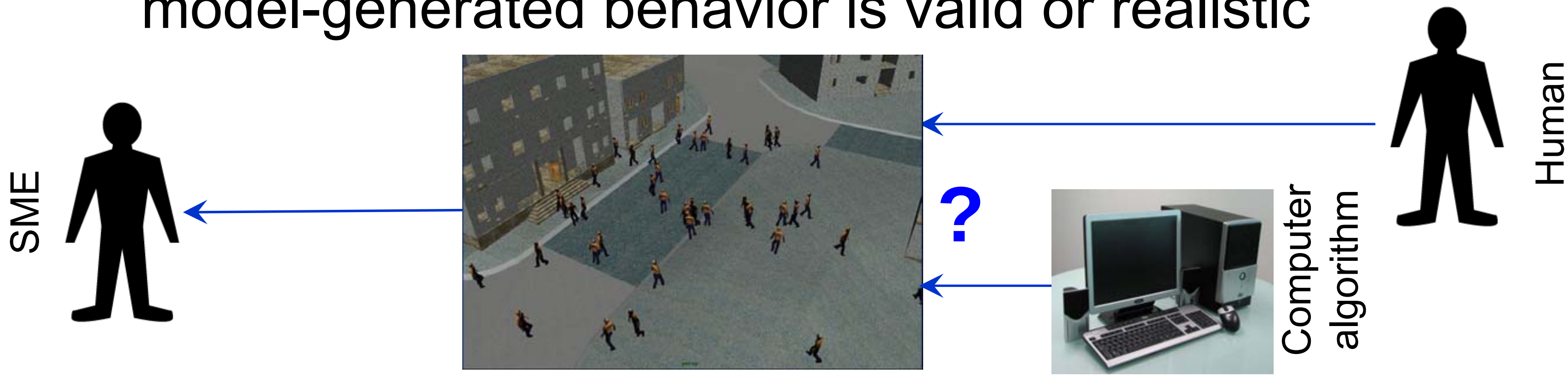




Turing test (validation) [Petty, 1994]

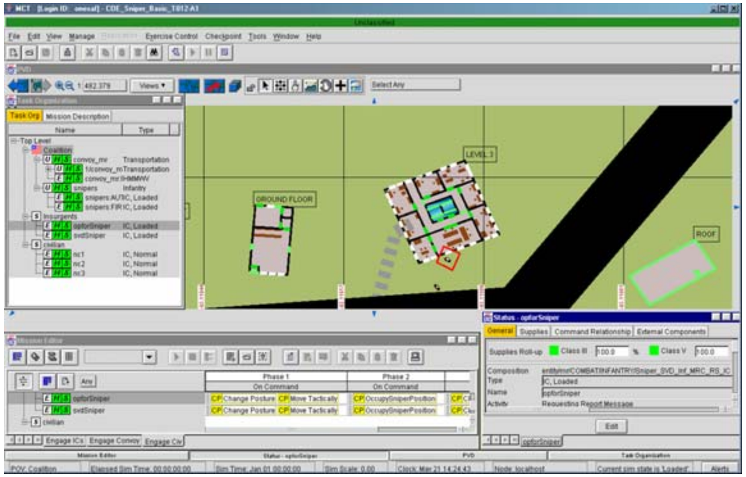
Turing

- Method
  - Subject Matter Experts observes behavior
  - Identify behavior as human- or model-generated
- Compares model behavior to human behavior
- Comments
  - Suitable primarily for human behavior models
  - Inability to reliably distinguish suggests model-generated behavior is valid or realistic



Semi-automated forces (SAF) systems

- Generate and control multiple simulated entities
- Used standalone or with other models
- Autonomous behavior for SAF entities
  - Generated by software in SAF model
  - Controlled by human operator via user interface
  - Military hierarchy represented



OneSAF

Turing test example [Potomac, 1990] [Wise, 1991]

- SIMNET (Simulator Networking)
  - Mounted combat team tactics training
  - Distributed, virtual, entity-level, real-time
  - Homogenous, proprietary
- SAF (Semi-Automated Forces)
  - Automated opponents within SIMNET
  - Behavior generated by software

M1 turret

M1 driver

Simulator bay

Out-the-window

- Experimental design
  - Soldiers in M1 simulators fought multiple tank battles
  - Two scenarios (1 and 2), two platoons (A and B)
  - Opponents were other platoon, SIMNET SAF, or both
- Results
  - Defenders not able to identify attackers, i.e., "pass"
  - Restricted field of view from simulators and small tactical behavior repertoire limited information

Experimental design

Scenario	Defender	Attacker
1	A	B
1	A	SAF
1	A	B + SAF
1	B	A
1	B	SAF
1	B	A + SAF

Scenario	Defender	Attacker
2	A	B
2	A	SAF
2	A	B + SAF
2	B	A
2	B	SAF
2	B	A + SAF

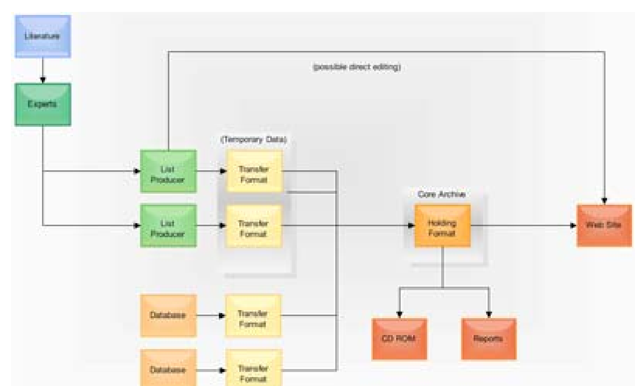
Static methods

Static V&V methods

- Characteristics
  - Methods based on artifact characteristics that can be determined without running a simulation
  - Often involve analysis of executable model code
  - May be supported by automated tools or manual notations or diagrams
  - More often performed by technical experts
- Example static V&V methods
  - Data analysis
  - Cause-effect graphing

Data analysis (verification)

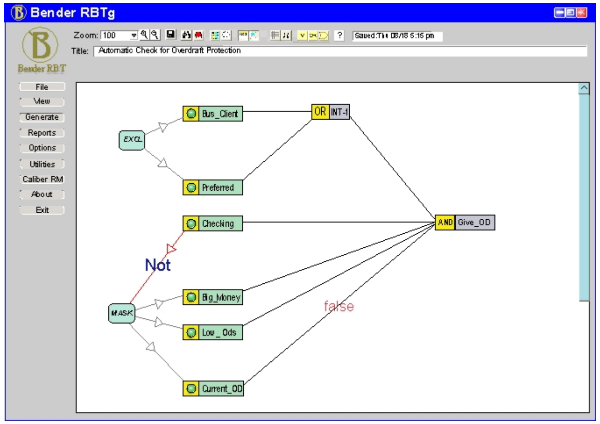
- Compare data definitions and operations in conceptual model to same in executable model
  - Data consistency
  - Data dependency analysis
  - Data flow analysis
- Compare conceptual model to executable model
- Determine if treatment and use of data consistent between artifacts



Data flow diagram

Cause-effect graphing (validation)

- Compare causes and effects in simuland to those in conceptual model
  - Cause: event or condition
  - Effect: state change triggered by cause
- Compare simuland to conceptual model
- Identify missing, extraneous, and inconsistent cause-effect relationships



Cause-effect graph

Dynamic methods

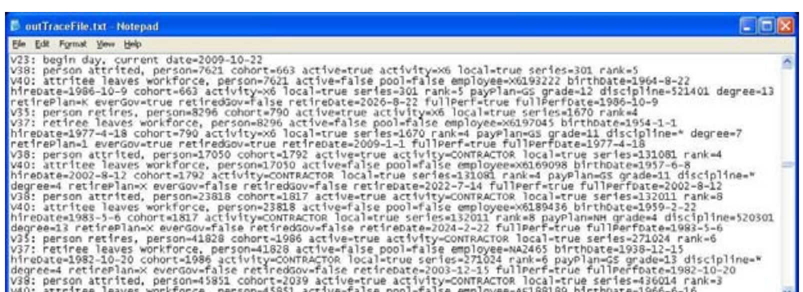
Dynamic V&V methods

- Characteristics
  - Methods that involve running the executable model and assessing the results
  - May compare results with simuland or other models
  - More often quantitative and objective
  - More often performed by technical experts
- Example dynamic V&V methods
  - Execution tracing
  - Sensitivity analysis
  - Comparison testing
  - Statistical methods

Execution tracing (verification or validation)

[Balci, 1998] [Mans, 2010]

- Record and examine simulation execution
  - "Line by line" or "step by step"
  - Output simulation state variables at each state change
  - Examine state for consistency, reasonableness
- Compare results to conceptual model, simuland
- Comments
  - Output may be to GUI or trace file
  - Examination may be manual or automated

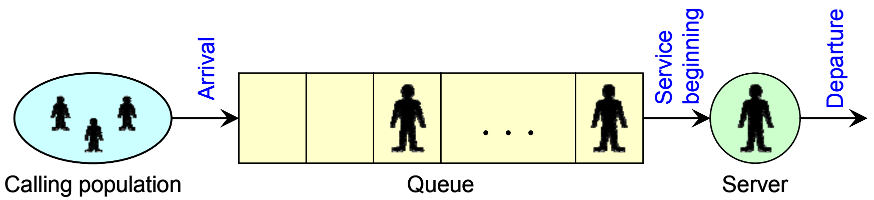


outTraceFile.txt - Notepad

File Edit Format View Help

```
V23: begin day, current date=2009-10-22
V38: person attrited, person=7621 cohort=663 active=true activity=X6 local=true series=301 rank=5
V40: attritee leaves workforce, person=7621 active=false pool=false employee=X6193222 birthDate=1964-8-22
hireDate=1986-10-9 cohort=663 activity=X6 local=true series=301 rank=5 payPlan=GS grade=12 discipline=521401 degree=13
retirePlan=K everGov=true retiredGov=false retireDate=2026-8-22 fullPerf=true fullPerfDate=1986-10-9
V35: person retires, person=8296 cohort=790 active=true activity=X6 local=true series=1670 rank=4
V37: retiree leaves workforce, person=8296 active=false pool=false employee=X6197045 birthDate=1954-1-1
hireDate=1977-4-18 cohort=790 activity=X6 local=true series=1670 rank=4 payPlan=GS grade=11 discipline=* degree=7
retirePlan=1 everGov=true retiredGov=true retireDate=2009-1-1 fullPerf=true fullPerfDate=1977-4-18
V38: person attrited, person=17050 cohort=1792 active=true activity=CONTRACTOR local=true series=131081 rank=4
V40: attritee leaves workforce, person=17050 active=false pool=false employee=X6169098 birthDate=1957-6-8
hireDate=2002-8-12 cohort=1792 activity=CONTRACTOR local=true series=131081 rank=4 payPlan=GS grade=11 discipline=*
degree=4 retirePlan=X everGov=false retiredGov=false retireDate=2022-7-14 fullPerf=true fullPerfDate=2002-8-12
V38: person attrited, person=23818 cohort=1817 active=true activity=CONTRACTOR local=true series=132011 rank=8
V40: attritee leaves workforce, person=23818 active=false pool=false employee=X6189436 birthDate=1959-2-22
hireDate=1983-5-6 cohort=1817 activity=CONTRACTOR local=true series=132011 rank=8 payPlan=NM grade=4 discipline=520301
degree=13 retirePlan=X everGov=false retiredGov=false retireDate=2024-2-22 fullPerf=true fullPerfDate=1983-5-6
V35: person retires, person=41828 cohort=1986 active=true activity=CONTRACTOR local=true series=271024 rank=6
V37: retiree leaves workforce, person=41828 active=false pool=false employee=NA2465 birthDate=1938-12-15
hireDate=1982-10-20 cohort=1986 activity=CONTRACTOR local=true series=271024 rank=6 payPlan=GS grade=13 discipline=*
degree=4 retirePlan=X everGov=false retiredGov=false retireDate=2003-12-15 fullPerf=true fullPerfDate=1982-10-20
V38: person attrited, person=45851 cohort=2039 active=true activity=CONTRACTOR local=true series=436014 rank=3
V40: attritee leaves workforce, person=45851 active=false pool=false employee=AE188189 birthDate=1966-6-16
```

Execution tracing example [Banks, 2010]



Simulation state variables

CLOCK	= Simulation time
EVTYP	= Event type (Start, Arrival, Departure, or Stop
NCUST	= Number of customers in queue at time given by CLOCK
STATUS	= Status of server (0 = Idle, 1 = Busy)

Event trace (status after event occurs)

CLOCK = 0	EVTYP = 'Start'	NCUST = 0	STATUS = 0
CLOCK = 3	EVTYP = 'Arrival'	NCUST = 1	STATUS = 0
CLOCK = 5	EVTYP = 'Depart'	NCUST = 0	STATUS = 0
CLOCK = 11	EVTYP = 'Arrival'	NCUST = 1	STATUS = 0
CLOCK = 12	EVTYP = 'Arrival'	NCUST = 2	STATUS = 1
CLOCK = 16	EVTYP = 'Depart'	NCUST = 1	STATUS = 1

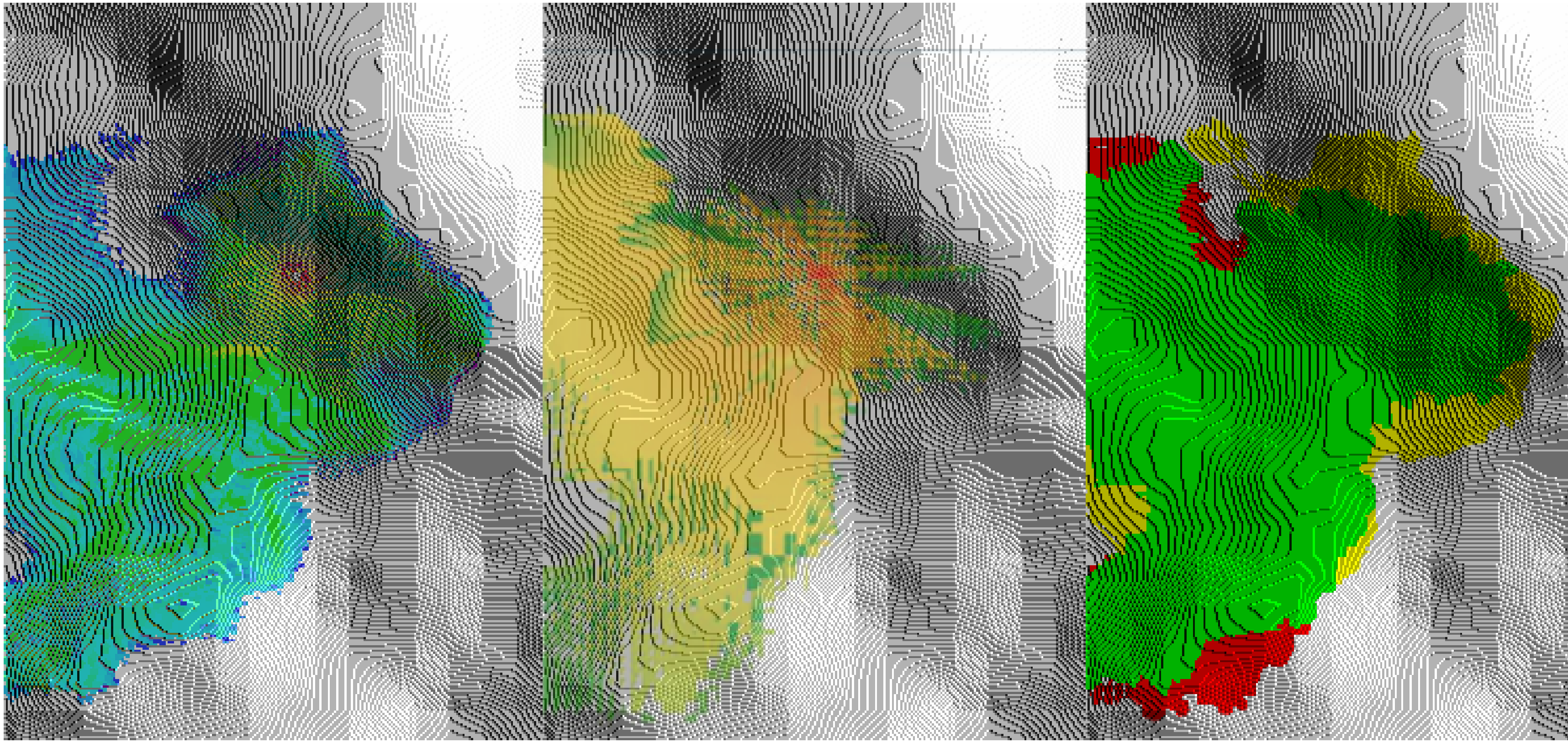
- Single server, single queue discrete event simulation
- Test produced mean queue length 0.4375, reasonable
- Trace reveals at time 3: queue length 1 and server status 0, error

Comparison testing (verification or validation)

- Run simulations of simuland (and scenario) using two different models, compare results
- Compare results to results
- Differences between results signal problems
- Comments
  - If differences, which model has problems?
  - If one model assumed valid, validation method
  - If neither model assumed valid, verification method

Comparison testing example [Filiposka, 2011]

- Durkin's radio propagation model
  - Estimates radio coverage area of a transmitter
  - Models attenuation caused by diffraction
  - Considers shadowing caused by terrain
  - Predicts transmission loss using path geometry
- Verified using comparison testing
  - Durkin's model compared to freely available Longley-Rice Irregular Terrain Model
  - Estimated radio coverage areas compared



Longley-Rice

Durkin's

Coverage comparison

Green = Both

Yellow = Longley-Rice only

Red = Durkin's only

[Filiposka, 2011]

Statistical methods (validation)

- Compare model results to simuland observations using statistical methods
  - Various statistical methods: regression analysis, analysis of variance, confidence intervals, hypothesis tests, others [Balci, 1998] [Petty, 2010]
  - May be used in combination with other methods
- Compare results to simuland
- Comments
  - Each statistical method defines statistic or metric of "closeness" or similarity; measure of validity
  - Generally underutilized

Example applications of statistical methods

Model(s)	Statistical method	Reason for selection
Spacecraft propulsion system sizing tool	Regression	Paired data, simuland–model
Medical clinic waiting Seaport loading/unloading Historical tank battle	Confidence intervals	Single simuland observation, multiple model runs
Bombing accuracy MC	Confidence intervals with error tolerance	Single simuland observation, multiple model runs, error tolerance available
Bank drive-up waiting line Entity-level combat	Hypothesis test comparing distributions	Multiple simuland observations, multiple model runs
Command decision making	Hypothesis test for equivalence	Multiple simuland observations, multiple model runs assumption of equality avoided
Missile impact MC	Hypothesis test comparing variances	Multiple simuland observations, multiple model runs

Formal methods

Formal V&V methods

- Characteristics
  - Methods based on formal mathematical proofs of program correctness
  - Quantitative (or logical) and objective
  - Performed by technical experts
  - Difficult to apply in practice [Balci, 1998]
- Example formal V&V methods
  - Inductive assertions
  - Predicate calculus

Inductive assertions (verification)

- Construct proof of executable model correctness
  - Assertions, statements about required executable model input-to-output relations, are associated with execution paths in executable model
  - Proofs of assertions along paths are constructed
  - Proofs along all paths imply correctness
- Compare executable model to conceptual model
- Comments
  - Closely related to general program proving techniques
  - Proofs done using mathematical induction
  - “Correctness” is with respect to conceptual model

Predicate calculus (validation)

- Logically analyze conceptual model
  - Predicate calculus is a formal logic system
  - Create, manipulate, and prove statements
  - Simuland, conceptual model described in pred calc
  - Prove properties of both to show logical consistence
- Compare conceptual model to simuland
- Quite difficult to apply to non-trivial problems

$$(\forall x)[D(x) \rightarrow (\forall y)(R(y) \rightarrow C(x, y))]$$
$$(\exists x)[D(x) \wedge (\forall y)(R(y) \rightarrow C(x, y))]$$
$$(\forall y)[R(y) \rightarrow (\forall x)(C(x, y) \rightarrow D(x))]$$
$$(\forall x)(\forall y)[R(y) \wedge C(x, y) \rightarrow D(x)]$$

Last two: "Only dogs chase rabbits." [Gersting, 2003]

Case study: Validation using confidence intervals

Confidence interval concept

- Basic terminology
  - Population; all “objects” of interest
  - Sample; selected subset of population
  - Parameter; numeric measure of population, e.g., mean
  - Statistic; numeric measure of sample, e.g., mean
- Confidence intervals as estimates
  - Sample mean point estimate of population mean
  - Range of values calculated from sample interval estimate of population mean
  - Calculated to have known confidence that population mean is within interval

Confidence interval formulas and critical values

- General form

  [*point estimate – margin of error*, *point estimate + margin of error*]

- Normal z distribution

$$\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}}\right]$$

- Student t distribution

$$\left[\bar{x} - t_c \frac{s}{\sqrt{n}}, \bar{x} + t_c \frac{s}{\sqrt{n}}\right]$$

Confidence level c	Normal z	Student t			
		d.f. = 5	d.f. = 10	d.f. = 20	d.f. = 30
0.80	1.282	1.476	1.372	1.325	1.310
0.90	1.645	2.015	1.812	1.725	1.697
0.95	1.960	2.571	2.228	2.086	2.042
0.99	2.576	4.032	3.169	2.845	2.750

Choosing a distribution

- Analyst must choose normal z or Student t
- Considerations
  - Population distribution: normal, approx normal, unknown
  - Population standard deviation σ: known, unknown
  - Sample size n: ≥ 30, < 30

If ((the population distribution is normal or approximately normal) or (the population distribution is unknown and the sample size $n \geq 30$)) and (the population standard deviation σ is known),
then calculate the confidence interval using z and σ.

If ((the population distribution is normal or approximately normal) or (the population distribution is unknown and the sample size $n \geq 30$)) and (the population standard deviation σ is unknown),
then calculate the confidence interval using t and s.

If (the population distribution is unknown and the sample size is < 30),
then a confidence interval can not be calculated.

These rules from [Brase, 2009]; sources differ.

Statistical interpretation

- Incorrect
  - "Confidence interval $[L, U]$ with confidence level c has a probability c of containing population mean μ"
  - L, U, μ all constants
  - Either $L \leq \mu \leq U$ or not; probability $= 0$ or 1
- Correct
  - "If many samples taken and confidence interval $[L, U]$ with confidence level c calculated for each sample, $(100 \cdot c)\%$ of them would contain population mean μ"

Validation method interpretation

- Population
  - All possible runs of model
  - Finite size on digital computer
- Model executions sample from population
- Confidence interval for model, not simuland
- Conventional validation interpretation
  - Simuland value within confidence interval implies model valid
  - No statistical justification or refutation

Validation procedure

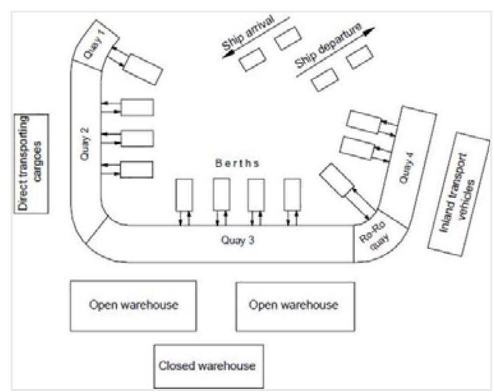
1. Select model response variable x to use for validation
2. Select number of model executions, i.e., sample size n
3. Execute model n times, producing sample $x_1, x_2, \ldots x_n$
4. Calculate sample mean $\bar{x}$ and sample std dev s
5. Select distribution normal z or Student t
6. Select confidence level c
7. Calculate confidence interval $[L, U]$
8. If known simuland value y within $[L, U]$, i.e., $L \leq y \leq U$, declare model valid (or not invalid) for x

Comments on validation procedure

- Assumes known simuland value y available
- Sample size $n \geq 30$ recommended
- Be cautious about assuming normality; recall that population is model, not simuland
- Confidence level $c = 0.95$ most common, some simulation experts recommend $c = 0.80$
- Simple inclusion test $L \leq y \leq U$ most common

Example: Seaport infrastructure [Demirci, 2003]

- Simuland
  - Seaport of Trabzon Turkey
  - Quays for berthing, unloading, loading ships
  - Three types of ships: G1, G2, G3
- Model
  - Discrete event simulation
  - Represents ships, cargos, quays, warehouses, cranes

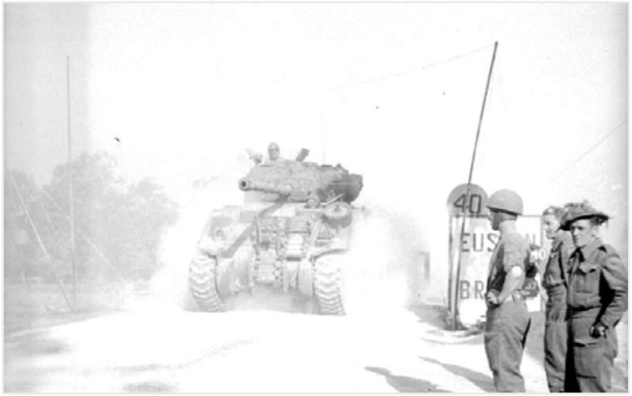


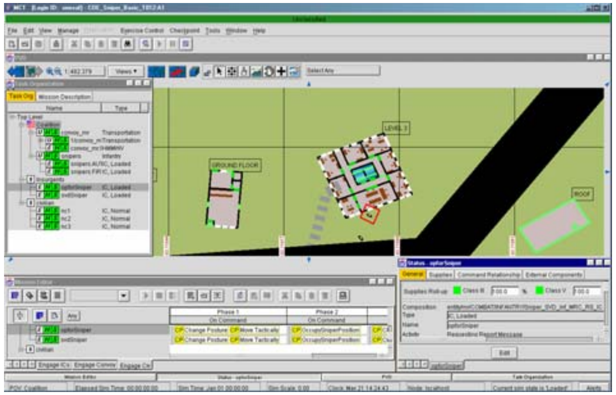
- Confidence intervals
  - Mean processed (count) for each ship type and total
  - Sample size (number of model runs) $n = 45$
  - Confidence level $c = 0.95$ (95%)
  - Distribution: Student t
  - Degrees of freedom $\mathrm{d.f.} = n - 1 = 44$
  - Critical value $t_c = 2.015$
- Results: 2 of 4 intervals contain simuland value

Ship Type	Simuland count	Model		Confidence interval		Within interval?
		Mean $\bar{x}$	Std dev s	L	U	
G1	109	111.14	14.45	106.8	115.5	Yes
G2	169	174.42	16.07	169.6	179.2	No
G3	19	17.28	5.26	15.7	18.8	No
Total	297	303.68	35.89	292.9	314.5	Yes

Example: WWII vehicle combat [Kelly, 2006]

- Simuland
  - Battle of Villers-Bocage, Normandy, June 1944
  - Small WWII tank battle, Britain vs Germany
  - Three types of British vehicles destroyed
- Model
  - OneSAF
  - WWII vehicle data (movement, P_k, P_h) added





- Confidence intervals
  - Mean destroyed (count) for each vehicle type
  - Sample size (number of model runs) $n = 20$
  - Confidence level $c = 0.95$ (95%)
  - Distribution: Student t
  - Degrees of freedom $\text{d.f.} = n - 1 = 19$
  - Critical value $t_c = 2.093$
- Results: 1 of 3 intervals contain simuland value

Vehicle Type	Simuland count	Model		Confidence interval		Within interval?
		Mean $\bar{x}$	Std dev s	L	U	
Firefly	4	1.6	0.502	1.365	1.835	No
Cromwell	10	5.3	1.695	4.510	6.093	No
Halftrack	10	9.2	2.745	7.915	10.485	Yes

Using this validation method

- Appropriate applications
  - Single simuland value for response variable
  - e.g., outcome of historical event
  - e.g., specific measurement
- Comments
  - If multiple simuland values available, alternate methods preferred (e.g., hypothesis test)
  - Historical outcome may have been atypical

Midway

73 Easting

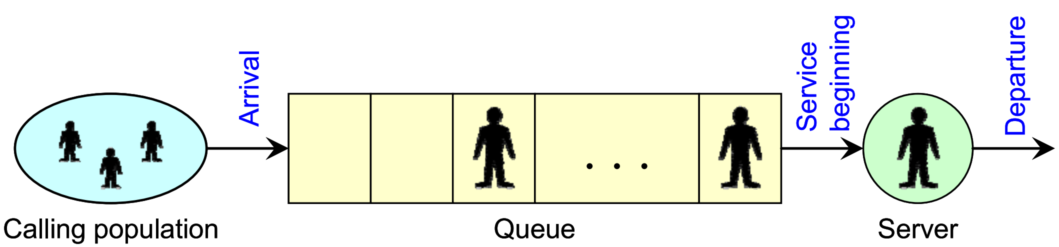
Case study summary

- Models
  - Seaport traffic; WWII combat
  - Different modeling paradigms
- Validation
  - Confidence intervals for means of model outputs
  - If confidence interval includes simuland value, model considered valid
- Lessons learned
  - Calculating confidence interval: easy
  - Determining suitable confidence level: not easy
  - Confidence interval useful when only one actual value available, e.g., historical result

Case study:
Validation using a statistical hypothesis test

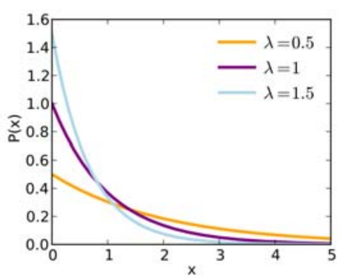
Fifth National Bank of Jaspar [Banks, 2010]

- Simuland
  - Bank drive-up window
  - Staffed by single teller; cars wait in single line
- Model
  - Conventional discrete event simulation
  - Single server, single queue
  - Simulate average delay (time spent in queue)

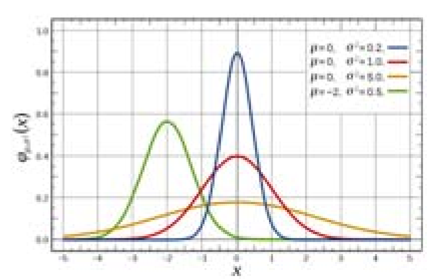


Data collection and modeling

- Simuland data collection
  - Collected for 90 customers, Friday 11:00am–1:00pm
  - Observed service times $S = \{S_1, S_2, \ldots, S_{90}\}$
  - Observed interarrival times $A = \{A_1, A_2, \ldots, A_{90}\}$
- Data modeling
  - Interarrival times: Exponentially distributed, rate $\lambda = 45$ per hour, mean $1/\lambda = 0.22$
  - Service times: Normally distributed, mean $\mu = 1.1$ minutes, standard deviation $\sigma = 0.2$ minutes



Exponential



Normal

Simuland and model variables

Variable names	Arrivals	Service times	Response
Simuland	A	S	Z
Model	W	X	Y

Decision variables

Number of servers: $D_1 = 1$
Number of queues: $D_2 = 1$

Input (stimulus) variables

Arrivals: $W_1, W_2, \ldots$
Service times: $X_1, X_2, \ldots$

Output (response) variables

Server utilization: Y_1
Mean delay: Y_2
Max queue length: Y_3
Arrival rate: Y_4
Mean service time: Y_5
Std dev service time: Y_6
Mean queue length: Y_7

Validation concept

- Mean delay important in queueing systems
- Compare model mean delay Y_2 from simulations to simuland mean delay Z_2 from observations
- Simuland $Z_2 = 4.3$ minutes (from observations)
- Comparison not simply comparing Z_2 and Y_2 and concluding "close enough"
- Hypothesis test statistically compares Z_2 and Y_2

Simulation results

Run	Y_4 Arrival rate	Y_5 Mean service time	Y_2 Mean delay
1	51	1.07	2.79
2	40	1.12	1.12
3	45.5	1.06	2.24
4	50.5	1.10	3.45
5	53	1.09	3.13
6	49	1.07	2.38
Sample mean $\bar{Y}_2$			2.51
Sample std dev s			0.82

$$\bar{Y}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_{2i} = 2.51 \text{ minutes}$$

$$s = \left[\frac{\sum_{i=1}^{n}(Y_{2i} - \bar{Y}_2)^2}{n-1}\right]^{1/2} = 0.82 \text{ minutes}$$

Statistical hypothesis test

- Student's t-test [Brase, 2009]
  - Determine if a sample is consistent with a population
  - Population (simuland) mean known, std dev unknown
  - Sample (model) mean known, std dev known
- Test structure
  - Hypotheses

    H_0: $E(Y_2) = 4.3$ minutes (model not invalid)

    H_1: $E(Y_2) \neq 4.3$ minutes (model invalid)
  - Level of significance $\alpha = 0.05$
  - Sample size $n = 6$

Critical value and test statistic

- Critical value of t [Brase, 2009]
  - Found in statistical table
  - Use $t_{\alpha/2,n-1}$ for two-sided test ($H_1 \neq$)
  - $t_{0.025,5} = 2.571$
- Test statistic
  - Quantifies discrepancy between sample mean and population mean
  - Compared to critical value

alpha one-tailed	.05	.025	.01	.005
alpha two-tailed	.10	.05	.02	.01
df				
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.743	4.604
5	2.015	$\boxed{2.571}$	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.869	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
inf	1.645	1.96	2.326	2.576

$$t_0 = \frac{\overline{Y}_2 - Z_2}{s/\sqrt{n}} = \frac{2.51 - 4.3}{0.82/\sqrt{6}} = -5.34$$

Test result and interpretation

- Rejection criteria for two-sided t test
  - If $|t_0| > t_{\alpha/2,n-1}$, then reject H_0
  - Otherwise, do not reject H_0
- Result
  - $|t_0| = 5.34 > t_{0.025,5} = 2.571$
  - Reject H_0
- Interpretation
  - Model is not valid w.r.t. mean delay
  - $P(H_0 \text{ rejected} \mid H_0 \text{ is true}) = \alpha = 0.05$ (Type I error)

V&V errors and statistical errors

	Model valid	Model not valid
Model used	Correct	**Type II error** Use of invalid model; Incorrect V&V; Model user's risk; **More** serious error
Model not used	**Type I error** Non-use of valid model; Insufficient V&V; Model builder's risk; **Less** serious error	Correct

Reject H_0 when H_0 true
i.e., reject a valid model
$P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(\text{Type I error}) = \alpha$

Fail to reject H_0 when H_1 true
i.e., fail to reject an invalid model
$P(\text{Fail to reject } H_0 \mid H_1 \text{ true}) = P(\text{Type II error}) = \beta$

Statistical power and validation

- Significance and power in statistical tests
  - Level of significance

    $P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(\text{Type I error}) = \alpha$
  - Power

    $1 - P(\text{Fail to reject } H_0 \mid H_1 \text{ true}) = 1 - P(\text{Type II error}) = 1 - \beta$
- Practical heuristics
  - To reduce $P(\text{Type I error})$, use small α
  - To reduce $P(\text{Type II error})$, use large n

Case study summary

- Model
  - Bank drive-up window
  - Conventional DES single server/single queue model
- Validation
  - Suitable statistical test (t-test) chosen for comparison
  - Population and sample means compared
- Lessons learned
  - Test revealed problem, opportunity to improve model
  - Rejecting H_0 stronger conclusion than not rejecting
  - Power can be increased with larger sample size

Case study:
Comparing real and simulated missile impact data

Introduction [Zhang, 2008]

- Application
  - Deterministic 6DOF model of missile trajectory
  - Used to calculate impact point given initial conditions
  - Measure x and y error w.r.t. aiming point
  - Compare model and live test x and y error variances
  - Two ranges: 60 Km and 100 Km
  - 6 live tests, 800 Monte Carlo model trials each range
- Monte Carlo analysis
  - For each trial, generate trajectory initial conditions from probability distributions
  - Calculate impact point
  - Repeat for 800 trials
  - Compare variances

Missile trajectory model

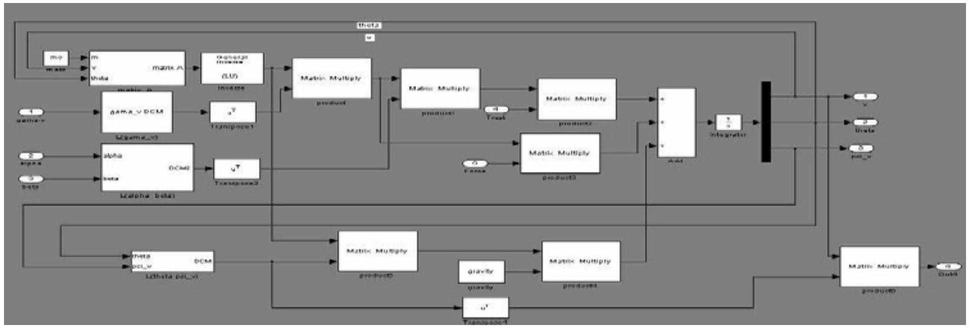
- Physics based
- Organized into modules: velocity, rotation, atmospheric conditions, aerodynamics, thrust
- Implemented in MATLAB Simulink

$$m\frac{dV}{dt} = P\cos\alpha\cos\beta - X - mg\sin\theta$$

$$mV\frac{d\theta}{dt} = p(\sin\alpha\cos\gamma_V + \cos\alpha\sin\beta\sin\gamma_V) + Y\cos\gamma_V - Z\sin\gamma_V - mg\cos\theta$$

$$-mV\cos\theta\frac{d\varphi_V}{dt} = P(\sin\alpha\sin\gamma_V - \cos\alpha\sin\beta\sin\gamma_V) + Y\sin\gamma_V + Z\cos\gamma_V$$

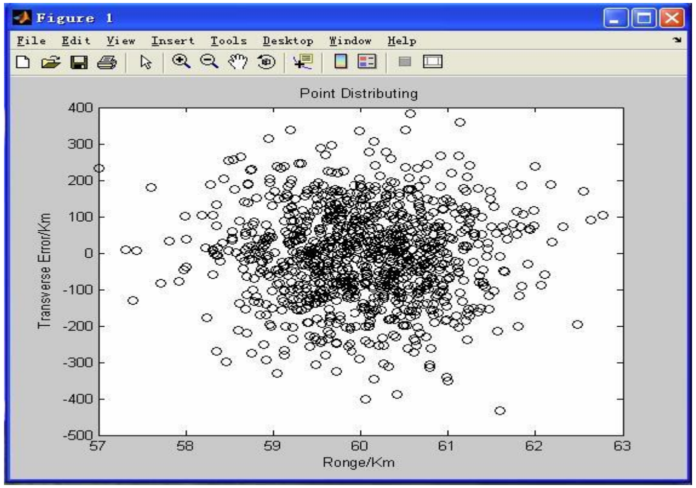
Velocity module equations



Velocity module block diagram

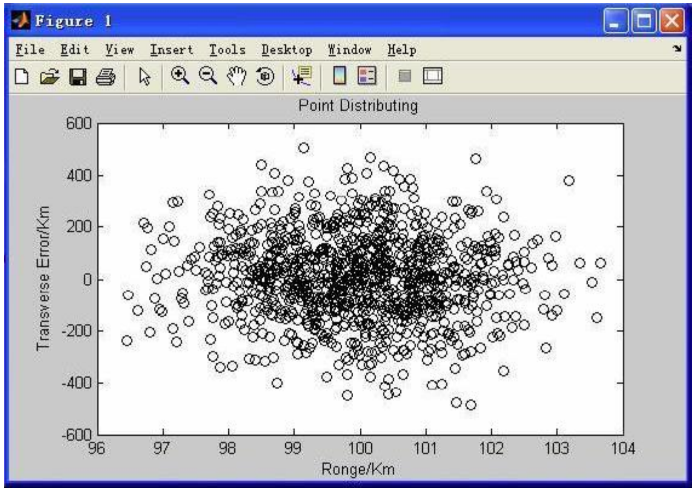
Impact data

60 Km



Trial	x error s	y error s	n
Model	526.62	85.91	800
Test	566.66	89.77	6

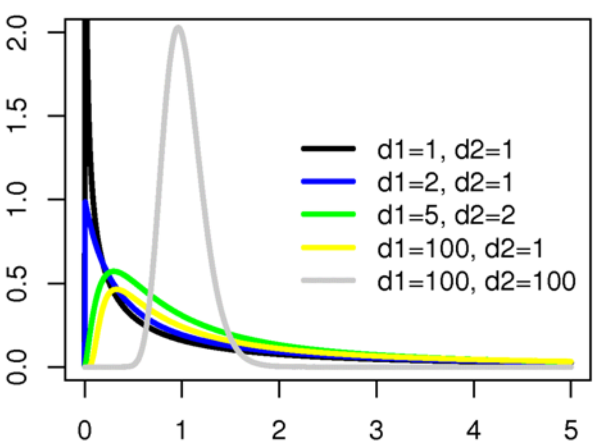
100 Km



Trial	x error s	y error s	n
Model	921.39	111.25	800
Test	980.52	120.68	6

Comparing variances: F test [Bhattacharyya, 1977]

- Compares variability of two populations
- Assumes both populations normally distributed
- Test statistic $F = s_1^2/s_2^2$
- Hypotheses (two-tailed test)
  - H_0: $\sigma_1^2/\sigma_2^2 = 1$ (variances equal)
  - H_1: $\sigma_1^2/\sigma_2^2 \neq 1$ (variances not equal)
- Reject H_0 if
  - If $F \geq F_{\alpha/2}(n_1-1, n_2-1)$ or
  - If $F \leq 1/F_{\alpha/2}(n_2-1, n_1-1)$



Applying the F test

- 60 Km missile impacts
- Test parameters
  - Level of significance $\alpha = 0.05$
  - Sample sizes $n_1 = 800$, $n_2 = 6$
- Critical values
  - $F_{\alpha/2}(n_1 - 1, n_2 - 1) = F_{0.025}(799, 5) = 6.0235$
  - $F_{\alpha/2}(n_2 - 1, n_1 - 1) = F_{0.025}(5, 799) = 2.5823$
  - $1/F_{\alpha/2}(n_2 - 1, n_1 - 1) = 1/F_{0.025}(5, 799) = 1/2.5823 = 0.3873$
- Test statistics and results
  - $F_x = 526.62^2/556.66^2 = 0.8950 < 6.0235$; do not reject H_0
  - $F_y = 85.91^2/89.77^2 = 0.9158 > 0.3873$; do not reject H_0

Comparing variances: Levene's test [Levene, 1960]

- Applicability of F test to missile impact data
  - Highly sensitive to assumption of normality
  - Potentially misleading results if populations not normal
- Levene's test
  - Compares variability of two populations
  - Does not assume populations normally distributed
- Test statistic $$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2}$$
- Variances not equal if $W \geq F_{\alpha}(k-1, N-k)$

Applying Levene's test

- 60 Km missile impacts
- Test parameters
  - Level of significance $\alpha = 0.05$
  - Sample sizes $n_1 = 800$, $n_2 = 6$, $N = n_1 + n_2 = 806$
  - Number of groups $k = 2$
- Critical value
  - $F_\alpha(k - 1, N - k) = F_{0.05}(1, 804) = 3.8531$
- Test statistics and results
  - $W_x = 0.0046$; do not reject H_0
  - $W_y = 24.6991$; reject H_0

Case study summary

- Model
  - Deterministic 6DOF model of missile trajectory
  - Used to calculate impact point given initial conditions
- Validation
  - Monte Carlo analysis, 800 trials and 6 live test
  - Model and simuland variances compared
- Lessons learned
  - Variances may be compared as well as means
  - Be attentive to hypothesis test assumptions

Summary

Tutorial summary

- Verification, validation, and accreditation address related but distinct questions
  - Verification: Was the model built right?
  - Validation: Was the right model built?
  - Accreditation: Is the model the right one for the job?
- Validity defined w.r.t. model's intended purpose
- VV&A involve comparisons
- Different types of risks are associated with VV&A
- Many VV&A methods available
- Statistics may be used for V&V comparisons

References

[Balci, 1981] O. Balci and R. G. Sargent, "A Methodology for cost-risk analysis in the statistical validation of simulation models", *Communications of the ACM*, Vol. 27, No. 4., pp. 190-197.

[Balci, 1985] O. Balci and R. E. Nance, "Formulated problem verification as an explicit requirement of model credibility", *Simulation*, Vol. 45, No. 2, pp. 76-86.

[Balci, 1998] O. Balci, "Verification, Validation, and Testing", in J. Banks (Editor), *Handbook of Simulation: Principles, Advances, Applications, and Practice*, John Wiley & Sons, New York NY, 1998, pp. 335-393.

[Banks, 2010] J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol, *Discrete-Event System Simulation, Fifth Edition*, Prentice Hall, Upper Saddle River NJ, 2010.

[Belfore, 2004] L. A. Belfore, J. J. Garcia, E. K. Lada, M. D. Petty, and W. P. Quinones, "Capabilities and Intended Uses of the Joint Operations Feasibility Tool", *Proceedings of the Spring 2004 Simulation Interoperability Workshop*, Arlington VA, April 18-23 2004, pp. 596-604.

[Bhattacharyya, 1977] G. K. Bhattacharyya and R. A. Johnson, *Statistical Concepts and Methods*, John Wiley & Sons, New York NY, 1977.

[Brase, 2009] C. H. Brase and C. P. Brase, *Understandable Statistics: Concepts and Methods*, Houghton Mifflin, Boston MA, 2009.

[Demirci, 2003] E. Demirci, "Simulation Modelling and Analysis of a Port Investment", *SIMULATION: Transactions of the Society for Modeling and Simulation International*, Vol. 79, Iss. 2, February 2003, pp. 94-105.

[DOD, 1996] Department of Defense, *Instruction 5000.61*, M&S VV&A, 1996.

[DOD, 2009] Department of Defense, *Instruction 5000.61*, M&S VV&A, 2009.

[Filiposka, 2011] S. Filiposka and D. Trajanov, "Terrain-aware three dimensional radio-propagation model extension for NS-2", *SIMULATION: Transactions of the Society for Modeling and Simulation International*, Volume 87, Numbers 1–2, January–February 2011, pp. 7-23.

[Gersting, 2003] J. L. Gersting, *Mathematical Structures for Computer Science, A Modern Treatment of Discrete Mathematics, Sixth Edition*, W. H. Freeman, New York NY, 2003.

[Kelly, 2006] K. M. Kelly, C. Finch, D. Tartaro, and S. Jaganathan, "Creating a World War II Combat Simulator Using OneSAF Objective System", *Proceedings of the 2006 Interservice/Industry Training, Simulation, and Education Conference*, Orlando FL, December 4-7 2006, pp. 510-520.

[Levene, 1960] H. Levene, "Robust tests for equality of variances", in I. Olkin, S. G Ghurge, W. Hoeffding, W. G. Madow, and H. B. Mann (Editors), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press, Palo Alto CA, 1960, pp. 278-292.

[Mans, 2010] R. S. Mans, N. C. Russell, W. van der Aalst, P. J. M. Bakker, and A. J. Moleman, "Simulation to Analyze the Impact of a Schedule-aware Workflow Management System", *SIMULATION: Transactions of the Society for Modeling and Simulation International*, Vol. 86, Iss. 8-9, August-September 2010, pp. 510-541.

[Petty, 1994] M. D. Petty, "The Turing Test as an Evaluation Criterion for Computer Generated Forces", *Proceedings of the Fourth Conference on Computer Generated Forces and Behavioral Representation*, Orlando FL, May 4-6 1994, pp. 107-116.

[Petty, 2009] M. D. Petty, "Verification and Validation", in J. A. Sokolowski and C. M. Banks (Editors), *Principles of Modeling and Simulation: A Multidisciplinary Approach*, John Wiley & Sons, Hoboken NJ, 2009, pp. 121-149.

[Petty, 2010] M. D. Petty, "Verification, Validation, and Accreditation", in J. A. Sokolowski and C. M. Banks (Editors), *Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains*, John Wiley & Sons, Hoboken NJ, 2010, pp. 325-372.

[Potomac, 1990] *Report of the Evaluation of the Representation of Semi-Automated Forces (SAF) in the SIMNET Model*, Potomac Systems Engineering, Annandale VA, 1990.

[Sargent, 1996] R. G. Sargent, "Verifying and Validating Simulation Models", *Proceedings of the 1996 Winter Simulation Conference*, Coronado CA, December 8-11 1996, pp. 55-64.

[Sargent, 2000] R. G. Sargent, "Verification, Validation, and Accreditation of Simulation Models", *Proceedings of the 2000 Winter Simulation Conference*, pp. 50-59.

[Shannon, 1975] R. E. Shannon, *Systems Simulation: The Art and Science*, Prentice Hall, Upper Saddle River NJ, 1975.

[Wise, 1991] B. P. Wise, D. Miller, and A. Z. Ceranowicz, “A Framework for Evaluating Computer Generated Forces”, *Proceedings of the Second Behavioral Representation and Computer Generated Forces Symposium*, Orlando FL, May 6-7 1991, pp. H1-H7.

[Zhang, 2008] W. Zhang, F. Li, Z. Wu, and R. Li, “Emulation of Rocket Trajectory Based on a Six Degree of Freedom Model”, *Proceedings of the Seventh International Symposium on Instrumentation and Control Technology: Measurement Theory and Systems and Aeronautical Equipment*, Proceedings of the SPIE Vol. 7128, Beijing China, October 10 2008, doi:10.1117/12.806878.

End notes

- More information
  - Mikel D. Petty, Ph.D.
  - University of Alabama in Huntsville
  - Center for Modeling, Simulation, and Analysis
  - 256-824-4368, pettym@uah.edu
- Questions?