8 Validierung von Modellen

Zur Erinnerung:

Modelle werden erstellt,

- um Experimente mit realen Systemen zu vermeiden
- um Aussagen über (potenzielle) Objekt-Systeme zu erhalten

Folgerungen aus der Modellanalyse
sollten weitestgehend gleich sein zu Folgerungen, die aus entsprechenden Objekt-Analysen/-Experimenten gewonnen würden

$\Rightarrow$ Es muss zumindest plausibel gemacht werden, dass gleiche/ähnliche Folgerungen aus Modell- und System-Experiment gezogen werden

Folgerungen sind dann identisch, wenn unmittelbare Beobachtungen/Resultate von Simulations- und Objekt-Experimenten identisch sind!

Ist Identität der Resultate zur erwarten?

Allgemein sicher nicht, da

- Objekt- und Modell-System sind nicht identisch,
- Objekt- und Modell-System zeigen stochastisches Verhalten

Verhaltensunterschiede sind zu erwarten!!

Zentrale Frage:

Welche Verhaltensunterschiede sind tolerierbar?

Grundlegende Definitionen

Benötigt wird eine vernünftige Definition von

Realitätstreue, Gültigkeit, validity

Aufbauend auf einem Maß für Verhaltensunterschiede

$D(V_R,V_S)$ (R: real, S: simuliert)

Eine solche Definition müsste

- Realitätstreue bejahen, falls Abweichungsmaß unter einem bestimmten Grenzwert liegt
- den Grenzwert so festlegen, dass (unvermeidlich existierende) Verhaltensunterschiede, die Folgerungen nicht beeinflussen

Eigentliche Situation:

- V_R nicht (vollständig) beobachtbar
  - auch dort wo V_R beobachtbar ist, treten Schwankungen auf
    $\Rightarrow$ nur eine Schätzung $\tilde{V}_R$ kann ermittelt werden
- V_S stammt aus einem stochastischen Modell
  $\Rightarrow$ nur eine Schätzung $\tilde{V}_S$ kann ermittelt werden
- Auf Basis einzelner Beobachtungen kann man $D(V_R, V_S)$ nicht direkt bestimmen, es gilt

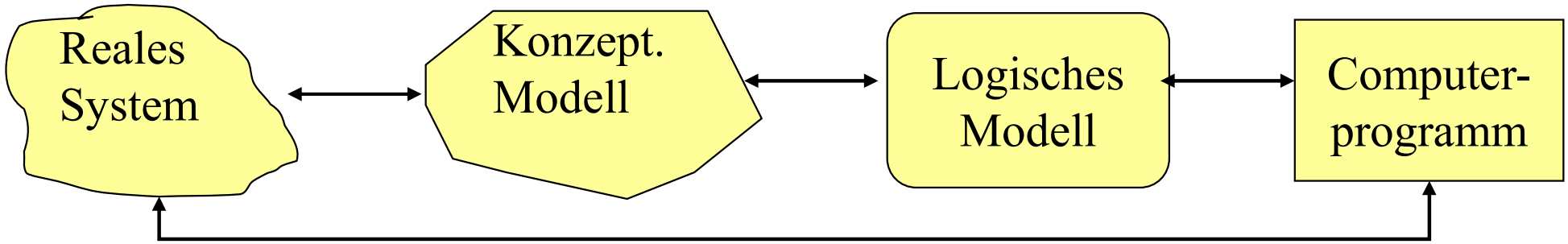
$$D(V_R, V_S) \leq \underbrace{D(\tilde{V}_R, V_R) + D(\tilde{V}_S, V_S)} + D(\tilde{V}_R, \tilde{V}_S)$$

Methoden siehe Kap. 5

Ursachen für Verhaltensunterschiede

- Strukturelle Unterschiede
  - Abstraktion, Aggregierung (beabsichtigt)
  - Ungenauigkeiten (bugs) (unbeabsichtigt)
  - Fehler (faults) (unbeabsichtigt)
- Parameterwerte
  - Deterministische Werte + Abhängigkeiten
  - Verteilungsfunktionen
- dazu stochastische Schwankungen

Transformationsschritte im Rahmen der Modellbildung:



Transformationsprozess von vage definierter Problemstellung (schlecht strukturiertem Realsystem)

zu wohl-definiertem Computerprogramm

Jeder angegebene Transformationsschritt kann zu neuen

Fehlern / Irrtümern / Verzerrungen

führen!

Oftmals uneinheitliche Terminologie, wir unterschieden hier:

- **Verifikation**
  Bestätigung aller Modell-Eingabegrößen/-Annahmen inkl. Struktureller Annahmen, Programmverifikation, Parameterwerte (mittels Testverfahren), …

- **Validierung**
  Bestätigung der Modellresultate

- **Kalibrierung**
  Reduktion der Verhaltensunterschiede
  (d.h. Reduktion von $D(V_R, V_S)$),
  falls Unterschiede als zu groß bewertet wurden
  Anpassung i.a. durch Änderungen am Modell

Auch wenn große Mühe auf eine „gute“ Wahl der Eingangsgrößen gelegt wurde

(d.h. Modellierung der Eingabegrößen, Testverfahren, …),

steigt zwar die „Hoffnung“ auf ein gültiges Modell,

eine „Garantie“ für ein gültiges Modell ist aber nicht gegeben!

Positivistischer Blick: Ergebnisse ok, alles ok

reicht oft nicht aus!

⇒ Kalibrierung ohne Verifikation / Validierung ist gefährlich!

In allen Schritten Beachtung des Kosten/Nutzen-Aspektes:

- Nutzen steigt mit dem Erkenntniswert
- Kosten steigen mit dem Aufwand der Erstellung

➢ Kompromiss zwischen Resultat und Aufwand ist notwendig

8.2 Verifikation

Oft Einsatz von (semi-)formalen/automatischen Techniken

Ein wichtiger Aspekt:

Verifikation des Simulationsprogramms am logischen/konzeptuellen Modelle (also Programmverifikation) hier nicht weiter behandelt!

Weitere Ansätze und Aspekte:

Zum Programmentwurf / zur Programmerstellung:

- Strukturierte Programmierung mit Test/Debugging von Modulen/Subprogrammen
- Code-Review durch andere Mitarbeiter
- Bei Verwendung von Spezifikationstechniken (semi-) automatische Generierung von Programmcode
- Inkrementeller Entwurf und Test

Simulationsspezifischere Ansätze:

- Test des Simulators für unterschiedliche Eingabeparameter (unterschiedliche Umgebungsbedingungen)
- Erstellen und analysieren von Traces
- Simulationsläufe unter vereinfachenden Annahmen (z.B. ohne ZVs), so dass Verhalten vorhersagbar
- „Durchspielen“ typischer Anwendungen
- Beobachtung von Animationen
- Verwendung von zuverlässigen Simulationswerkzeugen
- Verifikation der Eingabegrößen
  i.d.R. durch statistische Testverfahren
  (Verwendung des Begriffs Verifikation ist an dieser Stelle umstritten!)

8.3 Allgemeine Überlegungen zur Validierung

Simulationsmodell soll nützlich sein, d.h. mit sinnvoller Genauigkeit Aussagen über das System erlauben

$\Rightarrow$ **Es gibt keine absolute Validität von Simulatoren!**

Zu beachten ist

- Validierung ist modell-individuell
- Validierung ist graduell
- Validität ist oft Ergebnis eines (Verhandlungs-) Prozesses
- Validierung ist ein Projekt-begleitender Prozess

Man unterscheidet:

- Funktionsbezogene Validierung
  Test auf Plausibilität
- Ergebnisbezogene Validierung
  „Übereinstimmung“ der Ergebnisse System / Modell
- Theoriebezogene Validierung
  „Übereinstimmung“ der Ergebnisse analytisches / simulatives Modell

Wir betrachten im Folgenden primär die ergebnisbezogene Kalibrierung und Validierung !

Einsatz von Kalibrierung und Validierung setzt voraus,
dass V_R bestimmbar ist

- Daten resultieren aus Messungen am Realsystem
  - Daten sind oft „gestört“ oder müssen aufgearbeitet werden
  - Daten können nur in aggregierter Form oder für nicht relevante Situationen gewonnen werden
  - $\Rightarrow$ Probleme ähnlich zur Datenermittlung bei der Spezifikation quantitativer Modellgrößen
    (dort vorgestellte Ansätze auch hier verwenden)
- Daten resultieren aus anderen Systemen, anderen Modellen, Schätzungen, Vorhersagen
  - Im Prinzip ähnlich wie im ersten Fall, aber oft noch ungenauer

8.4 Schritte zur Kalibrierung von Modellen

- Identifikation der Ursachen von Verhaltensunterschieden
- entsprechende, gezielte Änderungen am Modell
  - Struktur, d.h. Code-Änderungen
  - Parameter, d.h. Werte-Änderungen

Unterstützung beim Finden der Ursachen:

- Strukturelle Ungenauigkeiten / Fehler sind meistens aus qualitativen Abläufen (Traces, Animationen) ableitbar
- Parameter Ungenauigkeiten / Fehler erfordern den Vergleich quantitativer Größen
  Hilfreich zur Eingrenzung:
  Vergleich von Zwischenergebnissen
  z.B. Verweilzeit an einer Station statt Gesamtverweilzeit

Bei stochastischen Modellen zwei spezifische Testprobleme

- Auswahl eines aus mehreren Modellen (S_1, S_2, …, S_K) auf Basis der zugehörigen $D(V_R, V_{Sk})$ $(k = 1,\ldots,K)$
  - D (wie gewohnt) Zufallsvariable
  - Unterschiede der D‘s müssen zur Auswahl einer Konfiguration „das normale Schwankungsmaß“ übersteigen
    $\Rightarrow$ Unterschiede müssen signifikant sein
    Aufgabe: **Test auf Signifikanz**

- „Tuning“ durch Parameterveränderungen, also Suche nach Parametervektor p_{opt} so dass
  $p_{opt} = \text{argmin}_p\, D(V_R, V_S(p))$
  Aufgabe: **Stochastische Optimierung**

Methodisch:

Kalibrieren eines stochastischen Simulators entspricht Experimentieren mit einem stochastischen Simulator

Also: Für gegebenes Realsystem mit Verhalten V_R finde Modell E mit Verhalten V_E, so dass $D(V_R, V_E)$ kleiner als vorgegebene Schranke

Möglichkeiten der Veränderung zur „besten/akzeptablen“ Alternative

- Ausprobieren struktureller Alternativen
- Tuning von System-Parametern

$\Rightarrow$ Methoden zum qualitativen und quantitativen Experimentieren mit stochastischen Systemen sollten hilfreich sein

Angenommen wir hätten D kleiner als erträgliches D_{ertr} erreicht

Sind dann Folgerungen aus Objekt- und Modell-Experiment identisch?

Simulator Verhalten wurde (mittels Änderungen) einem Objekt-Verhalten angepasst

- für einen Zustand der Umwelt
- für einen Zustand des Systems

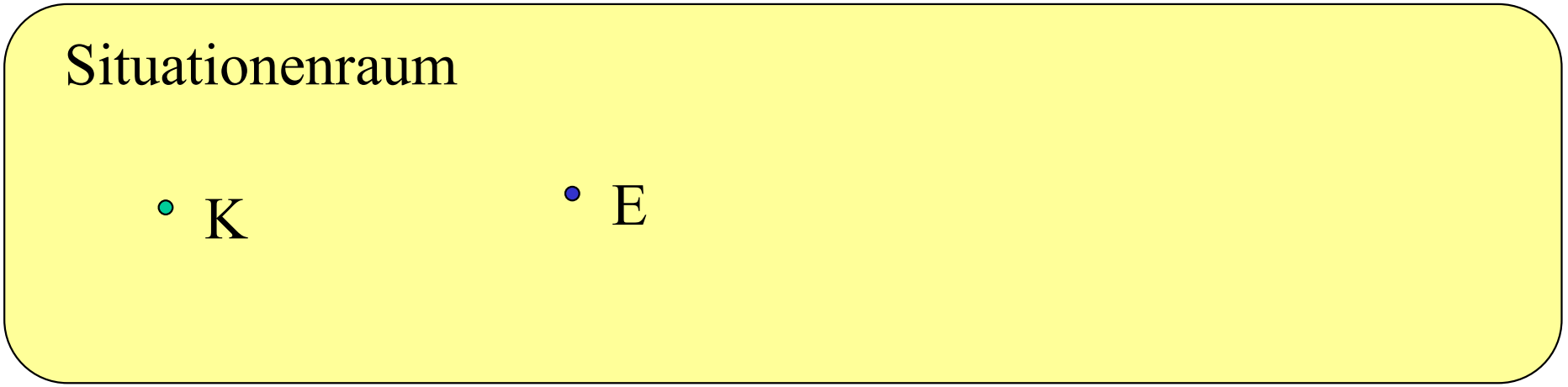
d.h. für genau eine Situation

Ziel der Simulation sind aber

- Aussagen über andere Situationen

  (das Resultat der Kalibriersituation ist ja bekannt)

Skizze des Prinzips



- Kalibrierung für Situation K
- Einsatz des Simulators für Situation E

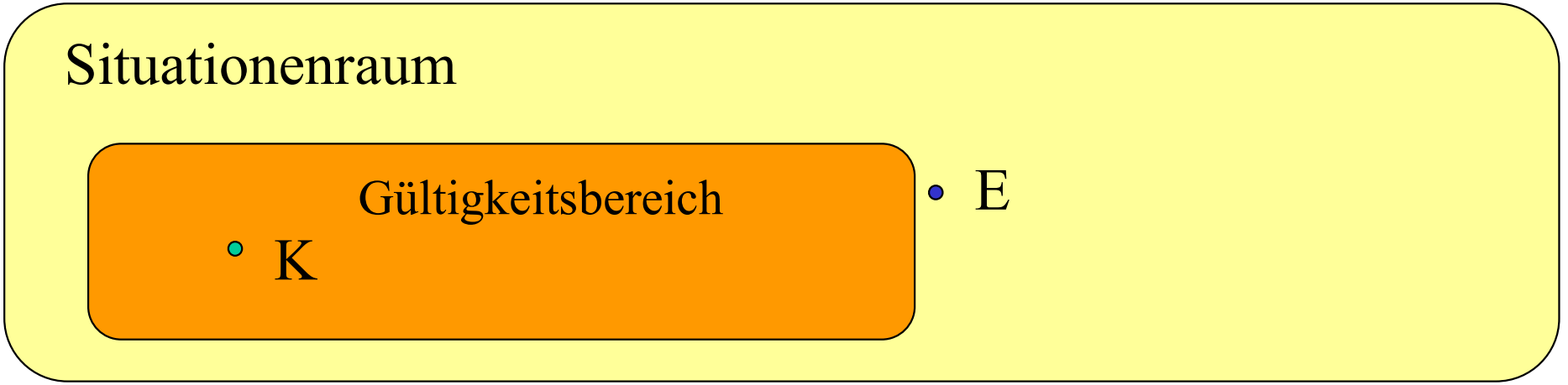
Zentrale Frage:

Ist der Unterschied D bei E gleich oder ähnlich dem Unterschied D bei K?

Dies ist u.U. fraglich!

Wenn Verhaltensunterschiede existieren, dann (wahrscheinlich) abhängig von der jeweiligen Situation, zu erwarten

- Bereich mit hinreichend kleinen Unterschieden: Gültigkeitsbereich G
- Bereich mit zu großen Unterschieden not G



Frage reduziert sich zu $E \in G$ oder $E \notin G$

Prinzipiell nur beantwortbar, wenn

- von Verhaltensunterschieden für gewisse Situationen
- auf Verhaltensunterschiede in anderen Situationen geschlossen werden kann

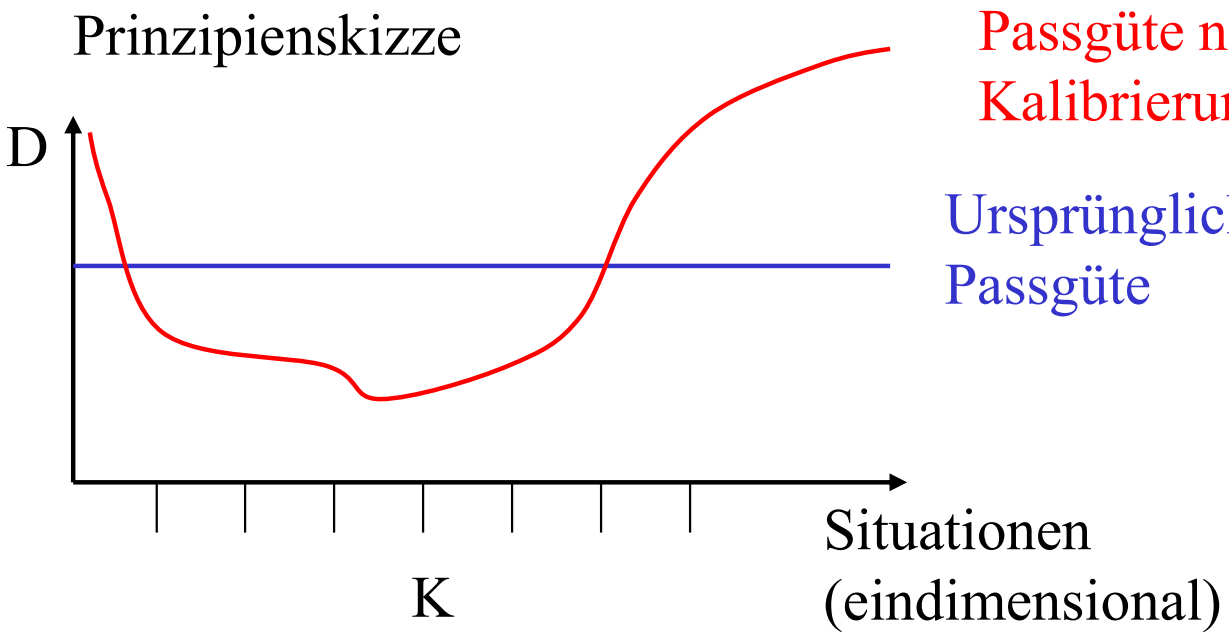
Dies gilt i.d.R. nicht, da dann Aussagen über das Systemverhalten für Situationen erforderlich sind, für die keine Experimente durchgeführt wurden

Damit (leider!) Übergang

- vom Beweis der Gültigkeit
- zur Zuversicht in Gültigkeit

bei jedem vorhersagendem Modellieren

Kalibriervorgang kann Problem durch Überanpassung noch verschärfen

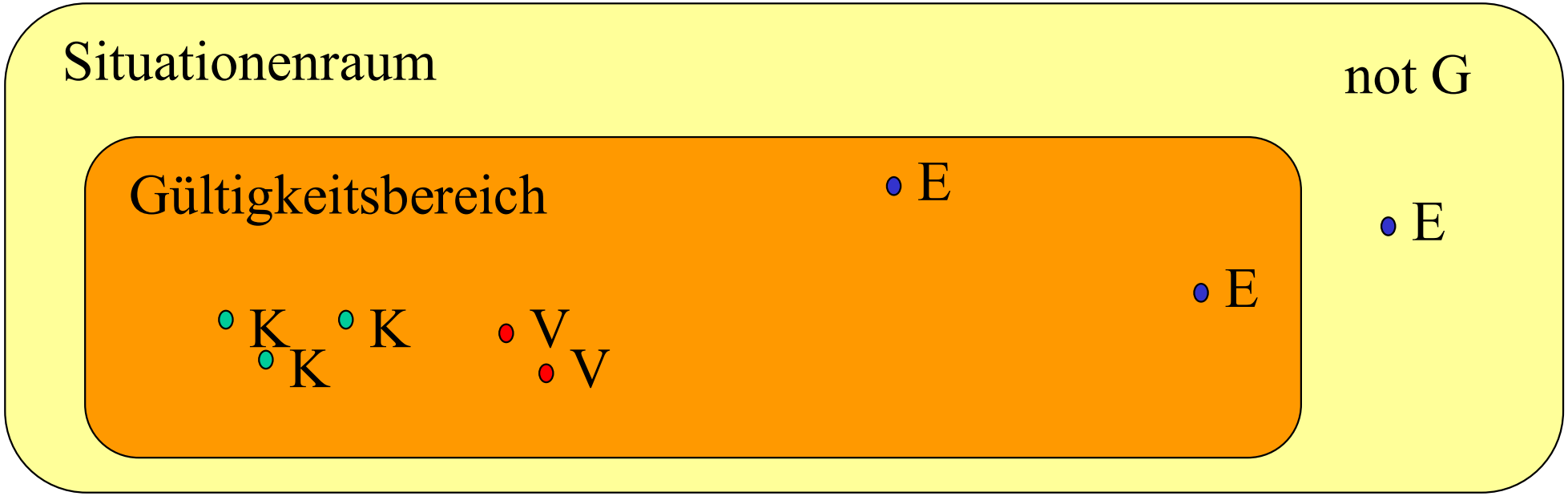


- Je nach Lage von E schlechtere Resultate nach der Kalibrierung
- Gefahr der Überanpassung kann reduziert werden, wenn für mehr als eine Situation kalibriert wird.

Dennoch ist formaler zu zeigen, dass Kalibrierung nicht zu Überanpassung führt!

Wie kann man plausibel machen, dass Model- und System-Verhalten hinreichend übereinstimmen?

- Spezielle und unabhängige Validierungsläufe für die Situationen, die nicht zur Kalibrierung genutzt wurden
- Prüfung, ob Verhaltensunterschiede für diese (Validier-) Situationen „in etwa so“ wie für bekannte (Kalibrier-) Situationen



K Kalibrier-Situation, V Validier-Situation, E Experiment-Situation

- Ob eine Experiment-Situation E im (unbekannten) Gültigkeitsbereich G liegt oder nicht, ist nach wie vor unbeweisbar (in irgendeinem strengen Sinne) bzw. ist nur retrospektiv beweisbar, dann aber uninteressant für Vorhersagen!
- Vertrauen in ein Modell wächst durch erfolgreiche Validierung (auch durch erfolgreiche retrospektive Validierung)
- Weitere Unterschiede auch bzgl. Abstand zu Validier- / Kalibrier-Situationen
  - Relativ große Sicherheit für Situationen E nahe bei bekannten Situationen K oder V
    (aber meistens weniger interessant)
  - Kaum Sicherheit für Situationen E weit weg von bekannten Situationen K oder V
    (aber oft interessant)

8.5 Messung von Verhaltensunterschieden

Bestimmung des Ausmaßes von Verhaltensunterschieden Objekt/Modell war wesentlich bei

- Kalibrierung: Realitätstreue des Modells verbessern durch Reduktion der Verhaltensunterschiede
- Validierung: Realitätstreue des Modells bestätigen durch Überprüfung von Verhaltensunterschieden in (von Kalibrierung) unabhängigen Situationen

Kalibrierung/Validierung soll/muss Vertrauen in hinreichende Realitätstreue unterstützen, denn Modell wird erstellt für Experimente

$\Rightarrow$ Modellgüte erhöhen durch Modellanpassung

Zentrale Fragestellungen beim „Messen von Verhaltensunterschieden“:

1. Mit dem Verhalten welchen Systems soll das Verhalten des Simulators verglichen werden?

2. Welche Verhaltens-Aspekte sollen für den Vergleich gewählt werden?

3. Welche Vergleichs-Methoden und –Techniken sollen eingesetzt werden?

Zu 1.

- Verhalten eines realen Objekts wäre ideal, aber
  - System muss existieren und beobachtbar sein
  - dann Verhaltensbeobachtung in verschiedenen Situationen
    - betreibe System und Modell in identischer Situation
    - oder verwenden vorhandener Aufzeichnungen über Systemverhalten, auch Aufzeichnungen aus „ähnlichen Systemen“
    - Trace-getriebene Simulation kann unterstützend wirken
    - nicht alle Daten für die Kalibrierung verwenden, sondern Daten für Validierung zurückhalten

Zu 1. (Fortsetzung)

- Verhalten analytischer Modelle
  - Analytisches Modell für Marginalsituationen entwickeln (z.B. ohne ZVs, bei sehr hoher Last, ..)
  - Ergebnisse der Simulation für diese Situationen mit analytischen Resultaten vergleichen
- Verhalten anderer Simulationsmodelle
  - Modelle ähnlicher Systeme
- Weitere Möglichkeiten (in der Praxis kaum verwendet)
  - Nutzung von Expertenwissen (z.B. kann Experte Ausgaben des Real-System von Ausgaben des Simulators unterscheiden?)

Zu 2.

- Ziel ist es, das System auf Basis bestimmter (ausgewählter) Leistungskriterien zu bewerten
  - Einsatz dieser Kriterien für den Vergleich
  - bei mehreren Kriterien oft Zielkonflikte
  - deshalb wird oft die Analyse auf ein Kriterium beschränkt, dann
    - Kalibrierung und Validierung bzgl. dieses Kriteriums
    - andere Kriterien nur zur Unterstützung
  - oder Einsatz mehrerer Modelle für die unterschiedlichen Kriterien

Zu 3. Fallunterscheidung bzgl. der eingesetzten Methoden

- Zum Vergleich Objektsystem – Simulator
  - Problemtyp: Vergleich zweier Stichproben
  - Entscheidung, ob zwei Stichproben hinreichend ähnlich
  - Oft beschränkt sich der Vergleich auf wenige Charakteristika (z.B. Mittelwerte)
  - Verfahren dazu auf den folgenden Folien
- Zum Vergleich analytisches Modell – Simulator
  - Problemtyp: Vergleich analytische Verteilung mit einer Stichprobe
  - Entscheidung, ob Stichprobe hinreichend ähnlich (d.h. aus analytischer Verteilung stammen könnte)
  - Verfahren kennen wir bereits (χ^2-Test, K.S.-Test, …)

Testverfahren

Ziel: Vergleich zweier Stichproben

(z.B. Verweilzeit Kunden am Bankschalter)

$v_R = (v_{R1}, \ldots\ldots , v_{Rn})$

$v_S = (v_{S1}, \ldots\ldots , v_{Sm})$

(z.B. R aus Beobachtungen Realsystem, S aus Simulation)

Man unterscheidet

- Subjektive Verfahren
  (basierend auf graphischen Darstellungen)
- Objektive Verfahren
  (basierend auf statistischen Tests oder dem Vergleich von Konfidenzintervallen)

Inspektionsansatz

Darstellung der Stichproben in graphischer Form durch

- Histogramme
- Punktdiagramme
- Box-Plots

Dargestellt werden

- alle Werte der Stichprobe oder
- eine zufällige Auswahl der Werte (bei Punktdiagrammen) oder
- die Stichprobe nach Elimination von Ausreißern

Anschließend Bewertung durch visuellen Vergleich

- Entwickler des Simulationsmodells
- Experten im Anwendungsbereich
- Turing Test

Probleme bei diesem Vorgehen

- Graphische Darstellung visualisiert nur einen Teil der Information
- Unterschiedliche Parametrisierung führt zu unterschiedlichen visuellen Eindrücken
- Keine „objektivierbaren“ Kriterien zur Entscheidungsfindung

Aber Inspektionsmethode ist ein erster wichtiger Schritt und kann mit statistischen Auswahlverfahren kombiniert werden

Vergleich von Konfidenzintervallen

Ziel: Aussage ob Mittelwerte gleich oder unterschiedlich
(deutlich restriktiver als bei der Inspektionsmethode!)

Voraussetzung:
Werte der beiden Stichproben seien unabhängig identisch verteilt!

Verschiedene Methoden existieren, im Wesentlichen Untersuchung, ob Konfidenzintervall für die Differenz der Mittelwerte 0 enthält

Wir betrachten hier zwei Verfahren

- Ungepaarte Stichproben (Welch-Verfahren)
- Gepaarte Stichproben (Paired t-Konfidenzintervalle)

Ungepaarte Stichproben (Welch Verfahren)

Zusätzliche Voraussetzung:

Werte zwischen den Stichproben sind unabhängig

aber keine Annahmen bzgl. identischer Varianz

Schätzer für die beiden Stichproben:

$$\tilde{\mu}_R = \frac{1}{n}\sum_{i=1}^{n} V_{R_i} \quad S_R^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(V_{R_i} - \tilde{\mu}_R\right)^2\right)$$

$$\tilde{\mu}_S = \frac{1}{m}\sum_{i=1}^{m} V_{S_i} \quad S_S^2 = \frac{1}{m-1}\left(\sum_{i=1}^{m}\left(V_{S_i} - \tilde{\mu}_S\right)^2\right)$$

Uns interessiert die Differenz $\mu_{RS} = \mu_R - \mu_S$

Für den allgemeinen Fall ist die Varianz nicht exakt zu ermitteln, folgende Approximation von Welch (1938) liefert aber i.d.R. gute Resultate

$\hat{\mu}_{RS} = \hat{\mu}_R - \hat{\mu}_S$ und $\hat{S}_{RS}^2 = {\hat{S}_R^2}/{n} + {\hat{S}_S^2}/{m}$

Schätzer für die Anzahl der Freiheitsgrade der –t-Verteilung:

$$\hat{f} = \frac{\left(\hat{S}_{RS}^2\right)^2}{\frac{1}{n-1}\left(\frac{\hat{S}_R^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{\hat{S}_S^2}{m}\right)^2}$$

Da $\hat{f}$ i.a. keine ganze Zahl, Wert runden oder Werte der Verteilung interpolieren!

Konfidenzintervall: $\hat{\mu}_{RS} \pm t_{\hat{f},1-\alpha/2} \cdot \hat{S}_{RS}$

Beispiel:

(aus Law/Kelton 00)

i	v_{R_i}	v_{S_i}	$v_{R_i} - v_{S_i}$
1	126.97	118.21	8.76
2	124.31	120.22	4.09
3	126.68	122.45	4.23
4	122.66	122.68	-0.02
5	127.23	119.40	7.83

Schätzwerte:

$\hat{\mu}_R = 125.57, \hat{\mu}_S = 120.59, \hat{S}_R^2 = 4.0$ und $\hat{S}_S^2 = 3.76$

sowie $\hat{\mu}_{RS} = 4.98, \hat{S}_{RS}^2 = 1.55\ und\ \hat{f} = 7.99.$

Konfidenzintervalle:

- [2.67, 7.29] zum Signifikanzniveau 0.1
- [0.81, 9.15] zum Signifikanzniveau 0.01

in beiden Fällen ist 0 nicht enthalten

Gepaarte Stichproben (Paired t-Konfidenzintervalle)

Zusätzliche Voraussetzung

n = m identische Beobachtungszahlen

aber keine Annahmen bzgl. identischer Varianz oder Unabhängigkeit der Stichproben!

Bestimme $\hat{\mu}_{RS}$ wie vorher und $\hat{S}_{RS}^2 = \frac{\sum_{i=1}^{n}\left(\left(v_{R_i} - v_{S_i}\right) - \hat{\mu}_{RS}\right)^2}{n(n-1)}$

Berechnung des Konfidenzintervalls: $\hat{\mu}_{RS} \pm t_{n-1,1-\alpha/2} \cdot \hat{S}_{RS}$

Beispiel:

Daten aus dem vorherigen Beispiel (siehe Folie 36)

Liefern $\hat{\mu}_{RS} = 4.98$ und $\hat{S}_{RS} = 1.56$

Konfidenzintervalle:

- [1.66, 8.31] zum Signifikanzniveau 0.1
- [-2.20, 12.16] zum Signifikanzniveau 0.01

Zum Signifikanzniveau 0.1 können die beiden Mittelwerte als ungleich angenommen werden, zum Signifikanzniveau 0.01 wird Gleichheit angenommen

Falls beide Verfahren anwendbar sind,

- ist nicht klar, welches schmalere Konfidenzintervalle liefert,
- sollte nach Datenlage entschieden werden

Weitere Methoden

Testverfahren

- Parametrisch (d.h. mit Verteilungsannahme) um $\mu_R \neq \mu_S$ zu testen
- Nicht-parametrisch (d.h. ohne Verteilungsannahme), oft wird auf $P[V_R < V_S] = 0.5$ getestet

Falls nur Mittelwerte verglichen werden sollen, sind Konfidenzintervalle zu bevorzugen, da sie quantitative Aussagen liefern, während Testverfahren eine binäre Antwort liefern.