7 Design of Experiments (LK 12)

7.1 Erkennen des Einflusses von Faktoren

- Das Systemverhalten (response) eines Simulationsmodells hängt von vielen Parametern (Modellfaktoren, factors) ab.
  - Beispiel: Wartezeiten bei einem M/M/1-Wartesystem hängen ab von der
    - mittleren Zwischenankunftszeit und von der
    - mittleren Bedienzeit.
  - Unterschiedliche Werte der Faktoren können für ein System realistisch und damit interessant zu untersuchen sein. Diese Wertebereiche sind für die Simulation von Interesse.
- Problem: Es ist in der Praxis oft nicht klar, welche Faktoren für das Systemverhalten wichtig sind. Das soll ja gerade herausgefunden werden.
- Ziel:
  - Ermittlung der wichtigsten Faktoren durch möglichst wenige Simulationsexperimente
  - Wie beeinflussen die Faktoren die Resultate?
  - Wie beeinflussen Kombinationen von Faktoren die Resultate?

7.1.1 Erschöpfende Simulation aller Faktoren

- Simulation aller Kombinationen aus den Faktorwertebereichen
- aus Rechenzeitgründen nicht möglich, da Anzahl der Simulation exponentiell mit der Anzahl der Faktoren steigt

7.1.2 Versuch & Irrtum Ansätze

- Faktorkombinationen werden zufällig gewählt und simuliert.
- Risiko zu hoch, entscheidende Faktoren zu übersehen.

7.1.3 Untersuchung von Einzelfaktoren

- Es gibt ein Referenzsystem
- Das System wird untersucht, indem für jeden Faktor einmal ein großer und ein kleiner Wert untersucht wird, wobei die restlichen Faktoren gleich bleiben
- Nur lineare Effekte erkennbar (r=c1*f1+c2*f2*…); Effekte auf Grund einer Kombination aus Faktoren nicht erkennbar (r=c12*f1*f2+c23*f1*f3).

7.1.4 2^k-Faktor-Designs (LK 12.2)

- Jeder Faktor hat nur zwei Werte (bezeichnet mit '+' und '-')
    - Bei k Faktoren sind dadurch 2^k Kombinationen der Faktorwerte möglich
    - Daher der Name: *„2^k factorial design"*
    - Jede dieser Kombinationen von k Faktoren ist ein *Designpunkt*
    - Ermittlung einer *Systemantwort* R_i für jeden Designpunkt i durch Simulationsläufe
    - Zusammenfassung der Designpunkte und Systemantworten in der *Designmatrix*

Beispiel: Designmatrix eines 2^3 factorial design:

Designpunkt	Faktor 1	Faktor 2	Faktor 3	Systemantwort
1	+	+	+	R_1
2	–	+	+	R_2
3	+	–	+	R_3
4	–	–	+	R_4
5	+	+	–	R_5
6	–	+	–	R_6
7	+	–	–	R_7
8	-	–	–	R_8

- Idee: Berechne: mittlere Änderung der Systemantwort bei der Änderung von Faktor j von '+' nach '–' ohne Änderung der anderen Faktoren
- Berechnung sogenannter Effekte von Faktorkombinationen (z.B. $x=f_1,f_2$)
- Beispiele
  - Berechnung des Effektes von Faktor f_1:

    $$e_1 = \frac{(R_1 - R_2) + (R_3 - R_4) + (R_5 - R_6) + (R_7 - R_8)}{4}$$
  - Berechnung des Effektes des Faktorkombination f_1, f_2:

    $$e_{12} = \frac{[(R_1 - R_2) - (R_3 - R_4)] + [(R_5 - R_6) - (R_7 - R_8)]}{4}$$
  - Berechnung des Effektes des Faktorkombination f_1, f_2, f_3:

    $$e_{123} = \frac{\{[(R_1 - R_2) - (R_3 - R_4)] - [(R_5 - R_6) - (R_7 - R_8)]\}}{4}$$
- Allgemeine Berechnung der Effekte für eine Faktorkombination: Summe über alle Systemantworten multipliziert mit den Vorzeichen (s_l, s_m, s_n) in der Designmatrix der entsprechenden Faktoren und geteilt durch 2^{k-1}:
  - $e_l = \frac{1}{2^{k-1}} \cdot \sum_{1 \le i \le 2^k} s_l(i) \cdot R(i)$
  - $e_{lm} = \frac{1}{2^{k-1}} \cdot \sum_{1 \le i \le 2^k} s_l(i) \cdot s_m(i) \cdot R(i)$
  - $e_{lmn} = \frac{1}{2^{k-1}} \cdot \sum_{1 \le i \le 2^k} s_l(i) \cdot s_m(i) \cdot s_n(i) \cdot R(i)$
- Interpretation

  Der Effekt einer Faktorkombination korreliert mit der Qualität und Stärke ihres Einflusses auf die Systemantwort
- Beispiel: Systemantwort

  $$R(f_1, f_2, f_3) = a_1 \cdot f_1 + a_2 \cdot f_2 + a_3 \cdot f_3 + a_{12} \cdot f_1 \cdot f_2 + a_{13} \cdot f_1 \cdot f_3 + a_{23} \cdot f_2 \cdot f_3 + a_{123} \cdot f_2 \cdot f_2 \cdot f_3$$

  In diesem Fall gilt sogar
  - $a_l = \frac{1}{2} e_l$
  - $a_{lm} = \frac{1}{2} e_{lm}$

- $a_{lmn} = \frac{1}{2} e_{lmn}$

Probleme:

- Voraussetzung linearer Abhängigkeit der Systemantwort von Faktorkombinationen
- Nur für beschränkte Anzahl von Faktoren anwendbar
- Einschränkung auf zwei Werte pro Faktor ⇒Einschränkung der Aussagekraft der beobachteten Effekte

Praktische Durchführung:

- Mehrfache Simulation $(n \geq 10)$ der Systemantworten für die 2^k Designpunkte der Designmatrix
- Erstellen von Konfidenzintervalle für die Systemantworten
- Einfluss einer Faktorkombination evident, falls das Konfidenzintervall der dazugehörigen Systemantwort die 0 nicht enthält.
- Aber: aus der statistischen Signifikanz eines Effekts folgt nicht notwendigerweise die praktische Signifikanz.

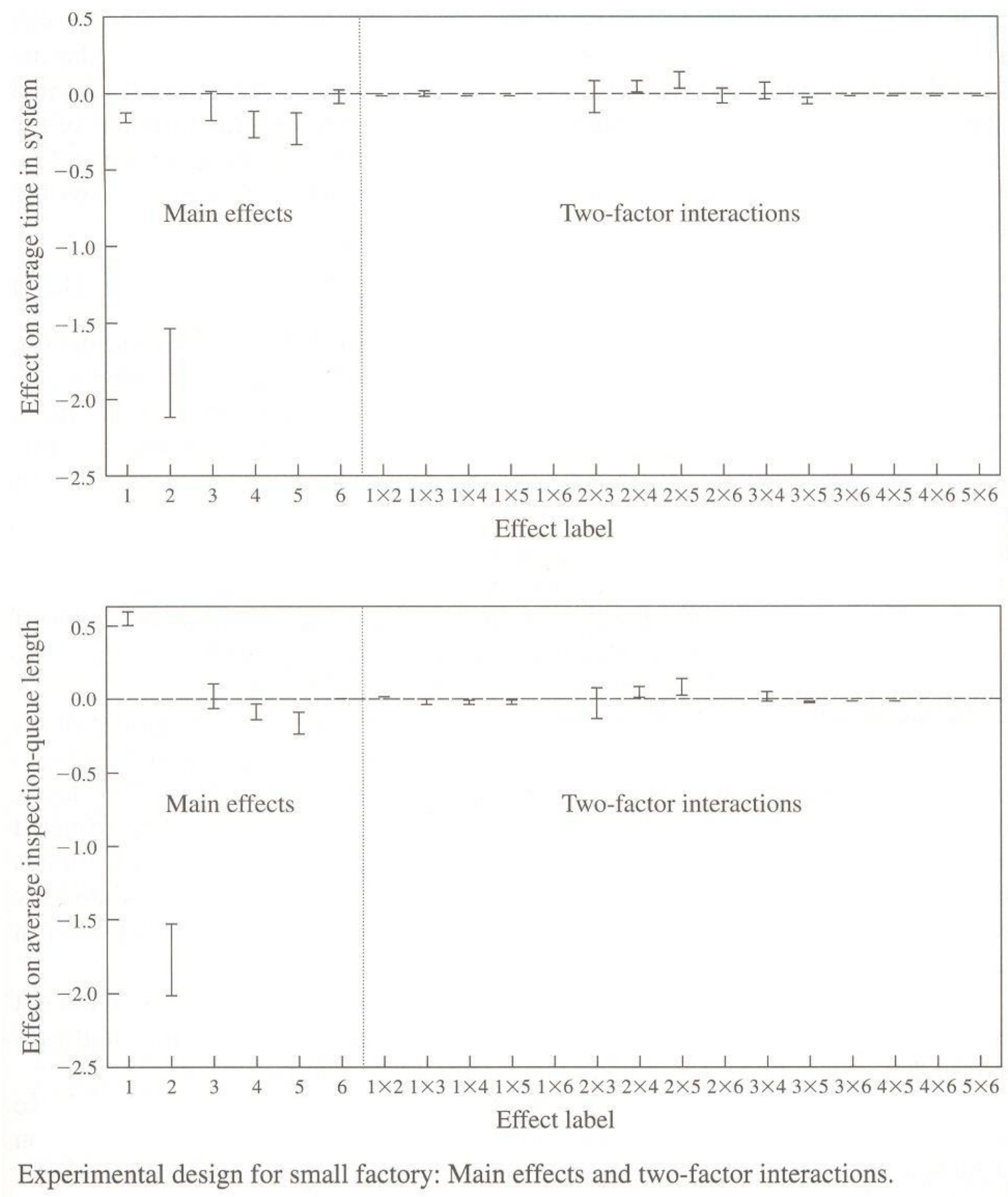

Abbildung 7.1: Ergebnisse eines Faktor Designs: Faktor 2 hat entscheidenden Einfluss auf die jeweilige Systemantwort, Faktoren 1, 4 und 5 haben geringen Einfluss, spezielle Faktorkombinationen haben keine Auswirkungen. (aus Law/Kelton: "Simulation Modeling and Analysis", 3rd Edition, S. 635)

<u>Lösung für große Anzahl von Faktoren:</u> 2^{k-p} *fractional factorial design*

- Nur 2^{k-p} Zeilen der kompletten Designmatrix werden untersucht
- Durch Streichung von Zeilen wird Aussagekraft verringert
- Herausforderung: geschicktes Streichen von Zeilen, so dass Effekte von einzelnen Faktoren sowie Effekte auf Grund von Zweierkombinationen von Faktoren noch erkennbar sind.
- Mehr dazu in LK 12.3.1

7.2 Systemverständnis und Reduktion von Parameteruntersuchungen

- System mit Eingangsparametern $\{x_1, \ldots, x_n\}$
- Durch Systemkenntnis: Zerlegung des Gesamtproblems
    - Erkennen der Auswirkung von Teileffekten in einem System (z.B. Verkehrsmenge) auf das Gesamtverhalten des Systems (z.B. Auslastung)
    - Erkennen von Parameterteilmengen, die den Teileffekt beeinflussen
    - Untersuchung der Auswirkung dieser Parameter auf den jeweiligen Teileffekt im System
    - Untersuchung des Einflusses der Teileffekte auf das Verhalten des Gesamtsystems indem z.B. nur der wichtigste Parameter eines Teileffektes verändert wird, Restparameter bleiben konstant ⇒ Reduktion von Parameteruntersuchungen.
    - Bei der Komposition von mehreren Teilsystemen: symmetrischer Versuchsaufbau zur besseren Kontrolle der Parameters (z.B. gleiche Ausfallwahrscheinlichkeit für alle Links eines Netzes)
- Erfassung und Untersuchung aller Systemparameter, z.B. auch die der Netztopologie
    - Netzgröße: Knoten und Linkanzahl
    - Vermaschungsgrad
    - Zusammenhangsstärke (wie viele Knoten kann man wegnehmen, ohne dass das Netz zerfällt)
    - Durchschnittliche Pfadlänge, Durchmesser
    - Zu welchen schon bekannten Teileffekten im System tragen sie bei?

- Beispiel: Paketverlustwahrscheinlichkeit auf einem Link beeinflusst durch
    - Linkkapazität c
    - Angebotene Last a in Erlang (Anzahl der angeschlossenen Teilnehmer, Teilnehmeraktivität, …)
    - Durchschnittliche Verbindungsrate r (Art der Applikation, unterschiedliche Benutzerklassen)
    - Variabilität der Verbindungen (Verkehrsarten, unterschiedliche Benutzerklassen)

  Durch Überlegung:
    - Anzahl der angeschlossenen Teilnehmer, Teilnehmeraktivität etc. beeinflussen angebotene Last (Teileffekt)
    - Auslastung $u=a*r/c$ wichtig für Leistungsbetrachtungen
    - Erkennen von Skalierungsparametern im System, z.B. c, a, r

- Die Paketverlustwahrscheinlichkeit liegt ausgelasteten Systemen in der Größenordnung $p = \frac{\max(a \cdot r - c, 0)}{a \cdot r}$
- Parameterstudien über a bei gleich bleibender Auslastung: Je größer die angebotene Last a ist, desto weniger schwankt das Gesamtverkehrsaggregat und desto kleiner ist p
- Variabilität des Gesamtverkehrsaggregats kann auch durch Variabilität der Verbindungen erhöht werden: erhöht auch die Paketverlustwahrscheinlichkeit p
- Untersuchung uninteressanter Parameterbereiche vermeiden, z.B. Parameterbereiche mit niedriger Auslastung führen zu 0% Paketverlusten

7.3 Quantifizieren und Darstellung des Einflusses von Faktoren

Faktoren haben i. A. nicht-linearen Einfluss auf die Systemantwort.
Problem: Wie können logarithmische, polynomielle, oder exponentielle Zusammenhänge erkannt und dargestellt werden?

Idee: Koordinatentransformation durch Skalierung der x- bzw. y-Achse
Platzierung der Koordinaten:

- Lineare Skalierung: $\hat{x}(x) = x$ bzw. $\hat{y}(y) = y$
- Logarithmische Skalierung: $\hat{x}(x) = \ln(x)$ bzw. $\hat{y}(y) = \ln(y)$
- Messpunkte sollen auch bei logarithmischen Skalierung der x-Achse in der Graphik gleichen Abstand voneinander haben, das erfordert einen exponentiellen Abstand der Eingangsparameter im Wertebereich.
- Beispiel:
  - Gewünscht: 10 Messwerte im Parameterbereich 1 bis 1000
  - Abstand der Messwerte im logarithmischen Bereich: $\frac{3-0}{10-1} = \frac{3}{9} = \frac{1}{3}$
  - Wähle Messwerte $x_k = 10^{\left(\frac{k}{3}\right)}$ mit $k \in [0;9]$

7.3.1 Skalierung zur Darstellung exponentiellen Wachstums

- Faktor x beeinflusst Systemantwort exponentiell: $y(x) = \exp(\lambda \cdot x)$
- Linear skalierte x-Achse: $\hat{x}(x) = x$
- Logarithmisch skalierte y-Achse: $\hat{y}(x) = \ln(y(x)) = \ln(\exp(\lambda \cdot x)) = \lambda \cdot x = \lambda \cdot \hat{x}(x)$
- Punkte im Koordinatensystem: $(\hat{x}(x), \lambda \cdot \hat{x}(x))$ ergeben eine Ursprungsgerade mit Steigung λ

7.3.2 Skalierung zur Darstellung logarithmischen Wachstums

- Faktor x beeinflusst Systemantwort logarithmisch: $y(x) = \ln(\lambda \cdot x)$
- Logarithmisch skalierte x-Achse: $\hat{x}(x) = \ln(x)$
- Linear skalierte y-Achse: $\hat{y}(x) = y(x) = \ln(\lambda \cdot x) = \ln(\lambda) + \ln(x) = \ln(\lambda) + \hat{x}(x)$
- Punkte im Koordinatensystem: $(\hat{x}(x), \ln(\lambda) + \hat{x}(x))$ ergeben eine Gerade, die die y-Achse bei $\ln(\lambda)$ schneidet

7.3.3 Skalierung zur Darstellung polynomiellen Wachstums

- Faktor x beeinflusst Systemantwort polynomiell: $y(x) = x^{\lambda}$
- Logarithmisch skalierte x-Achse: $\hat{x}(x) = \ln(x)$
- Logarithmisch skalierte y-Achse: $\hat{y}(x) = \ln(y(x)) = \ln\left(x^{\lambda}\right) = \lambda \cdot \ln(x) = \lambda \cdot \hat{x}(x)$

- Punkte im Koordinatensystem: $(\hat{x}(x), \lambda \cdot \hat{x}(x))$ ergeben eine Ursprungsgerade mit Steigung λ

7.3.4 Skalierung zur Darstellung linearen Wachstums

- Faktor x beeinflusst Systemantwort polynomiell: $y(x) = \lambda \cdot x$
- Linear skalierte x- und y-Achse: $\hat{x}(x) = x$, $\hat{y}(x) = y(x) = \lambda \cdot x = \lambda \cdot \hat{x}(x)$
  - Punkte im Koordinatensystem: $(\hat{x}(x), \lambda \cdot \hat{x}(x))$ ergeben eine Ursprungsgerade mit Steigung λ.
  - Evtl. Abweichungen im Bereich kleiner Parameter ist kaum sichtbar, da bei äquidistantem Abtasten der x-Achse nur große Parameter betrachtet werden.
- Logarithmisch skalierte x- und y-Achse: $\hat{x}(x) = \ln(x)$, $\hat{y}(x) = \ln(y(x)) = \ln(\lambda \cdot x) = \ln(\lambda) \cdot \ln(x) = \ln(\lambda) \cdot \hat{x}(x)$
  - Punkte im Koordinatensystem: $(\hat{x}(x), \ln(\lambda) \cdot \hat{x}(x))$ ergeben eine Ursprungsgerade mit Steigung $\ln(\lambda)$.
  - Evtl. Abweichungen im Bereich kleiner Parameter sind gut sichtbar, da alle Größenordnungen bei äquidistantem Abtasten der x-Achse untersucht werden.

7.3.5 Organisation der Ergebnisse in PivotTables zur flexiblen Darstellung

Problembeschreibung

- Man hat es oft mit sehr vielen Systemfaktoren zu tun, z.B. mit x1, x2, x3, x4, … und hat für deren Kombinationen Systemantworten ermittelt.
- Die Daten liegen oft in folgender Form vor:

x1	x2	x3	R(x1,x2,x3)
1	4	6	24
1	4	7	28
1	5	6	30
1	5	7	35
2	4	6	48
2	4	7	56
2	5	6	60
2	5	7	70
3	4	6	72
3	4	7	84
3	5	6	90
3	5	7	105

Ergebnistabelle

- Darstellungsmöglichkeiten für die Systemantwort
  - Zweidimensionales Diagramm: R(x1, x2, x3) auf y-Achse, ein Faktor auf x-Achse, z.B. x1, ein anderer, z.B. x2, wird als Parameter für eine Kurvenschar benutzt.
  - Dreidimensionales Diagramm: R(x1, x2, x3, …) auf y-Achse, zwei Faktoren, z.B. x1 und x2, werden auf den zwei x-Achsen dargestellt.
- Für die nicht berücksichtigten Faktoren kann dabei
  - entweder ein fester Wert angenommen werden, z.B. x3=6,
  - oder die Systemantwort wird über alle werte von x3, also x3=6 und x3=7, gemittelt.
- Problem: oft ist der Haupteinflussfaktor für die Systemantwort a priori noch nicht bekannt und muss noch gefunden werden.
  - Probiere x1, x2, und x3 jeweils als x-Achse für die Darstellung aus!
- Randbedingung vieler Darstellungsprogramme wie z.B. Excel
  - x- und y-Werte müssen als Vektoren gegeben sein, indem sie in einer Spalte oder Zeile aufeinander folgen.
  - Problem: das ist in der obigen Ergebniszusammenstellung nur für x3 mit festem x1 und x2 gegeben. Beispiel: x3=[6,7] mit R(1,4,[6,7])=[24,28].
  - Für andere Darstellungen muss die Ergebnistabelle umorganisiert werden, das übernimmt der Pivot-Table.

<u>Benutzung einer Pivot-Tabelle in Excel</u>

- Organisation der Daten wie in der Ergebnistabelle nebenan
  - Oberste Zeile: Parameterbezeichnungen
  - Restliche Zeilen: Ergebniswerte
- Durchführung des PivotTable-Assistenten (Daten->Pivot-Tabelle)
  - Einlesen der gesamten Tabelle inklusive Überschrift
  - Konfiguration des PivotTables
    - Es erscheint eine PivotTable-Feldliste
    - Ziehe den Faktor für die x-Achse in das Zeilenfeld
    - Ziehe den Faktor für die Kurvenschaar in das Spaltenfeld
    - Ziehe die restlichen Faktoren in die Seitenfelder
    - Ziehe die Systemantwort in das Datenfeld
    - Doppelklicke in das linke obere Eck des PivotTables, wo sich Spalten- und Zeilenfelder kreuzen und stelle von Summe auf Mittelwert um!

x3	(Alle)

Mittelwert von R(x1,x2,x3)	x2		
x1	4	5	Gesamtergebnis
1	26	32,5	29,25
2	52	65	58,5
3	78	97,5	87,75
Gesamtergebnis	52	65	58,5

Pivot-Tabelle

- In den Datenfeldern stehen keine Einzelwerte sondern die Durchschnitte über alle sonstigen Faktoren in den Seitenfeldern, man kann aber auch auf einen bestimmten Wert umstellen.
- Parameter im Seitenfeld, Zeilenfeld, Spaltenfeld können beliebig verschoben werden und somit ist es einfach Vektoren im Datenfeld gegen unterschiedliche Vektoren im Zeilenfeld zu plotten.
- Zum Plotten selber: kopiere die Vektoren aus dem PivotTable und erstelle mit hilfe des Diagramm-Assistenten eine XY-Graphik! Wähle aus, ob Spalten oder Zeilen gegeneinander geplottet werden sollen.