# Dynamische Verkehrs- und Preismodellierung für den Einsatz in Kommunikationssystemen

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
genehmigte Dissertation

vorgelegt von

Diplom–Mathematiker Univ.
Peter Reichl

aus Pappenheim

Berichter: Prof. Dr. Otto Spaniol
Prof. Dr. Boudewijn Haverkort

Tag der mündlichen Prüfung: 16. Dezember 1999

D82 (Diss. RWTH Aachen)

# *Zum Geleit, oder: “Was wissen die von Worten oder Werten ...”*

Dieses Büchlein ist erwachsen aus der Arbeit von vier wechselvollen Jahren, die mich von Aachen über New Jersey nach Zürich geführt haben. Vielen Menschen bin ich auf diesem Weg begegnet, und beinahe jeder (und jede) hat auf seine Weise einen Beitrag dazu geliefert. Ich bin froh, ihnen allen wenigstens auf diesem Weg von ganzem Herzen Danke sagen zu können.

Doch wo beginnen? Vielleicht wirklich am Anfang des Weges, bei Prof. Otto Spaniol vom Lehrstuhl i4 der RWTH Aachen. Die Freiheit, die er der Kreativität an seinem Lehrstuhl läßt, ist ja bereits Legende (vgl. obiges Zitat aus [Neb81]), und ich bin dankbar dafür, daß ich sie weidlich nutzen durfte. An ihn geht vor allem auch mein Dank für die Betreuung dieser Arbeit, ebenso wie an Prof. Boudewijn Haverkort für das Korreferat – mit ihm zu arbeiten hat mir immer große Freude und Genugtuung bereitet.

Stellvertretend für alle Aachener Mitstreiter sei “Kollesche Schuba” genannt, der nicht nur drei Jahre lang heldenhaft und unerschütterlich ein Büro mit mir teilte und mir bei meinen zahllosen “unlösbaren Problemen” zur Seite stand, sondern mir am Schluß auch noch den schönsten Doktorhut der Weltgeschichte gebastelt hat. Danke Claudia Linnhoff-Popien für die viele Unterstützung und Aufmerksamkeit, die mir immer wieder von so großer Hilfe war. Danke all den Helens, Rainers und Ulrichs, die mir unermüdlich bei der Bewältigung des Umgangs mit dem Rechner geholfen haben. Danke Axel, Olaf, Tony, Jens, Matz, Maria, Raschid, Christian, Henrik, Dirk, Dogan, Roland, Bernd, Simon, Andreas und all den anderen i4-lern für unzählige Gespräche, Antworten, Anregungen, Gesten. Und danke Petra, Höppi, Kerstin und Frau Gelück für die viele Mühe, die ebenso unersetzlich für eine reibungslose Arbeit ist wie sie allzuoft unbemerkt bleibt.

Sehr froh bin ich aber auch über die stimulierende Wirkung, die mein Aufenthalt bei Bell Laboratories (Murray Hill) in der Gruppe von Debasis Mitra und im Anschluß daran mein Wechsel in die Gruppe von Prof. Bernhard Plattner am TIK der ETH Zürich hatte. An beiden Orten wurde ich mit offenen Armen aufgenommen und habe mich auch in der Folge sehr wohl gefühlt. Danke an Burkhard Stiller für die vielen fachlichen und persönlichen Impulse und das wirklich allzeit offene Ohr, danke George, Nathalie, David und allen, die mir im Rahmen des CATI-Projekts zur Seite gestanden sind. Christina Class hat mir nicht nur das Eingewöhnen in die die TIK-Umgebung sehr einfach gemacht. Marcus, Ulrich, Marcel und Marcel, Rolf, Thomas, Hans-Jörg, Catarina, Susi – auch diese Liste kann nicht vollständig sein.

Von großer Hilfe waren mir aber auch die vielen Diskussionen und Anregungen, die ich im Austausch mit “meinen” Studenten bzw. Diplomanden erhielt. Tatjana Traikovska, Frank Mohren und Beat Spielmann haben mir durch das fachliche wie persönliche Engagement, mit dem sie ihre Diplomarbeiten angefertigt haben, eine kaum zu ersetzende Unterstützung zukommen lassen. Und dann gibt es noch die vielen, die zwar nichts direkt mit der vorliegenden Arbeit zu tun hatten, aber ohne die sie niemals zustande gekommen wäre. Stellvertretend für alle diese sei Udo Krieger von der Deutschen Telekom genannt, der in drei entscheidenden Augenblicken an mich geglaubt hat. Bettina Stich ist mit mir während dreier Jahre durch viele Höhen und Tiefen gegangen und war mir immer wieder eine große seelische Stütze. Der letzte Dank gebührt schließlich meinen Eltern und meiner ganzen Familie für all die Opfer, die sie gebracht haben, um mir meinen Weg überhaupt erst möglich zu machen.

Zürich, im Dezember 1999

Peter Reichl

# Inhaltsverzeichnis

**Kapitel 1 Einleitung – oder: ”... dynamisch, praktisch, gut.” 1**

**Kapitel 2 Dynamische Verkehrsmodellierung: Einführung und Problemstellung 5**

2.1 Mobilkommunikation unter GSM . . . . . 5
Übersicht über den GSM-Standard . . . . . 5
GSM-Systemtest . . . . . 7
2.2 Das Referenzproblem: Dynamische Modellierung von Location Updates . . . . . 8
2.3 Wichtige Verfahren der Verkehrsmodellierung . . . . . 10
Erneuerungsprozesse . . . . . 11
Markovprozesse . . . . . 12
Flußmodelle . . . . . 13
Autoregressive Modelle . . . . . 13

**Kapitel 3 Klassische Autoregressive Zeitreihenmodellierung 15**

3.1 Grundlagen autoregressiver Modelle . . . . . 15
Stationarität und Filter . . . . . 15
AR(p)-Prozesse . . . . . 18
ARMA(p,q)-Prozesse . . . . . 20
Ansatz von Box-Jenkins . . . . . 23
3.2 Modellierung mit AR-Prozessen . . . . . 25
Modellidentifikation . . . . . 26
Parametrisierungsverfahren . . . . . 27
Beispiel: Ungefilterte Modellierung der Referenzreihe . . . . . 30
3.3 Modellierung mit ARMA-Prozessen . . . . . 33
Filterung . . . . . 33
Modellidentifikation . . . . . 35
Verfahren zur Parameterschätzung . . . . . 35
Resultat und Diagnose . . . . . 37
3.4 Fazit . . . . . 38

**Kapitel 4 Die TES-Methode und ihre Verallgemeinerung 39**

4.1 Grundlagen und Standardverfahren . . . . . 39
Anforderungen an das TES-Verfahren: die “Drei Kriterien” . . . . . 39
Die Ebenen des TES-Schemas . . . . . 40
Die Autokorrelationsfunktion . . . . . 43
4.2 Parametrisierung und Resultate . . . . . 46
4.3 Generalized Stitching Function: Idee, Konsequenzen, Ergebnisse. . . . . 49
Der Einfluß der Stitching-Funktion und die Idee einer Verallgemeinerung . . . . . 49
Die Bestimmung einer Verallgemeinerten Stitchingfunktion . . . . . 51
Konsequenzen für die Distortion-Transformation . . . . . 52
Ergebnisse für die Referenzkurve . . . . . 56
Weitere Anwendung: ein Generalized TES-Modell zur Modellierung von MPEG-codiertem Videoverkehr . . . . . 59
Die verfeinerte GSF im Referenzmodell . . . . . 63
4.4 Automatisierung von TES . . . . . 65
Grundsätzliche Überlegungen und Vorgehensweise . . . . . 65
Der Automatisierungsalgorithmus . . . . . 66
Ergebnisse . . . . . 68
4.5 Fazit . . . . . 69

**Kapitel 5 Ergänzende Bemerkungen zur TES-Modellierung 71**

5.1 Offene Fragen . . . . . 71
Iterative Automatisierung von TES . . . . . 71
Formale Behandlung des Kriteriums der visuellen Ähnlichkeit . . . . . 72
Auflösungsqualität des Modells . . . . . 74
5.2 Ausblick: Ein TES-Modell für selbstähnlichen Verkehr?. . . . . 75

**Kapitel 6 Preismodelle für Kommunikationssysteme: Übersicht und Herausforderungen 79**

6.1 Überblick und Klassifizierung . . . . . 79
Ein Blick in die Praxis . . . . . 79
Zur Klassifizierung dynamischer Preismodelle . . . . . 80
6.2 Ansätze und Probleme . . . . . 83
Edge Pricing . . . . . 83
Auktionsmechanismen . . . . . 84
Profile und Klassen . . . . . 86
6.3 Anforderungen aus der Sicht eines Charging-und-Accounting-Tools . . . . . 87
IntServ und DiffServ . . . . . 88
RSVP . . . . . 90
Preismodelle in Multiprovider-Szenarien . . . . . 93

**Kapitel 7 Verallgemeinerte Preisfunktionen in Stochastischen Verlustnetzen 95**

7.1 Modell und Preisfunktionen . . . . . **95**
Kelly's Bound . . . . . **95**
Modellbeschreibung . . . . . **96**
Ableitung der Preisfunktion . . . . . **99**
Einschränkungen und Erweiterungsansätze . . . . . **101**
7.2 Approximation mit UAA und RUAA . . . . . **102**
Zur Darstellung der Partitionsfunktion als Kreisintegral . . . . . **102**
Grundidee der Approximation . . . . . **104**
Algorithmische Darstellung . . . . . **106**
Zusammenfassung . . . . . **109**
7.3 Validierung und Resultate . . . . . **109**
Der Referenzfall . . . . . **109**
Große Kapazitäten . . . . . **111**
Mischklassen-Szenarien . . . . . **112**
7.4 Preisfunktionen im Mehrklassenfall . . . . . **114**
Mehrklassen-Szenario als Folge von Mischklassen-Szenarien . . . . . **114**
Ein Beispiel . . . . . **116**
Zur Abhängigkeit von der momentanen Kapazitätsaufteilung . . . . . **118**
7.5 Zusammenfassung, Interpretation und Ausblick . . . . . **120**

**Kapitel 8 Auktionsbasierte Preismodelle für Multiprovider-Szenarien 123**

8.1 Einführung . . . . . **123**
8.2 Auktionsmechanismen und Utility-Funktion . . . . . **124**
Formale Definition einer Auktion . . . . . **124**
Nutzerpräferenzen, Utility-Funktion und Incentive Compatibility . . . . . **125**
Exkurs: Nutzerpräferenzen im Trader eines Verteilten Systems . . . . . **126**
Klassische Auktionsmechanismen . . . . . **126**
8.3 Anforderungen im Multiprovider-Szenario . . . . . **128**
8.4 Delta-Auktionen . . . . . **129**
8.5 CHiPS: Das Connection-Holder-is-Preferred-Scheme . . . . . **131**
Grundzüge des Verfahrens . . . . . **131**
Auktionen unter CHiPS . . . . . **133**
Konsequenzen . . . . . **134**
Implementationsaspekte . . . . . **135**
8.6 Simulation von Multiprovider-Szenarien unter FlowSim . . . . . **139**
Die FlowSim-Umgebung . . . . . **139**
Festlegung des Simulationsszenarios . . . . . **140**
Simulationsergebnisse . . . . . **141**
8.7 Fazit . . . . . **144**

**Kapitel 9 Ergänzende Bemerkungen zur Modellierung von Internet-Tarifen 145**

9.1 Zur Benutzerakzeptanz dynamischer Preismodelle . . . . . **145**

Das INDEX-Experiment . . . . . **145**

Ein Parameter zur Begrenzung von Preisschwankungen . . . . . **146**

9.2 Auf dem Weg zum Edge Pricing - oder: “Effektive Bandbreite” einmal anders **147**

9.3 Charging-und-Accounting-Technologie im Internet: Das Projekt CATI . . . . . **149**

**Kapitel 10 Schlußbemerkung 151**

**Anhang A Mathematische Grundlagen und Ergänzungen 153**

A.1 Zum mathematischen Hintergrund von TES . . . . . **153**

A.2 Analytische Berechnung der Autokorrelationsfunktion eines einfachen TES-Modells . . . . . **158**

A.3 Beispiel für den Automatisierungsalgorithmus . . . . . **164**

A.4 Erweiterung der Refined Uniform Asymptotic Approximation . . . . . **167**

**Anhang B Zur Berücksichtigung von Nutzerpräferenzen bei der Dienstvermittlung in einem Trader 169**

B.1 Einführung . . . . . **169**

B.2 Trading im Kontext von CORBA . . . . . **170**

B.3 QoS- und QoSP-Funktionen . . . . . **172**

B.4 Modellierung von Nutzerinteressen . . . . . **173**

B.5 Implementierung und Ergebnisse . . . . . **177**

B.6 Fazit . . . . . **178**

**Anhang C Abkürzungsverzeichnis 179**

**Anhang D Abbildungs- und Tabellenverzeichnis 181**

**Anhang E Literaturverzeichnis 185**

KAPITEL 1

# *Einleitung – oder: ”... dynamisch, praktisch, gut.”*

Hand in Hand mit der rasant wachsenden Bedeutung elektronischer Kommunikationstechnologie hat in den letzten Jahren auch die Modellierung der zugehörigen Systeme einen unerwarteten Aufschwung genommen. Während nun auf der theoretisch orientierten Seite einiges an Fortschritten zu verzeichnen ist, gestaltet sich die unmittelbare Verwendung dieser Ergebnisse in praktischen Anwendungen oft schwieriger als erwartet. Ein kurzer Blick auf industrienahe Projekte zeigt, daß die hier verwendeten Modelle in vielen Fällen noch relativ einfach sind – insbesondere finden sich oft *statische Modelle*, obwohl die untersuchten Fragestellungen in Wirklichkeit eine signifikante Zeitabhängigkeit aufweisen. Der auf diese Weise erzielte Gewinn in Sachen Komplexität und Rechenzeit wird dabei auf Kosten der Realitätsnähe erkauft. *Dynamische Modelle* sind zwar in der Regel aufwendiger zu implementieren und handzuhaben, bilden aber dafür die Wirklichkeit deutlich besser ab und erlauben daher eine fundiertere Interpretation der Modellierungsresultate. Vor diesem Hintergrund stellt sich immer wieder die dringende Frage, ob es gelingt, durch Anpassung etablierter Ansätze oder Verwendung neuer Ideen dynamische Modellierungstechniken bereitzustellen, die einfach genug sind, um in der Praxis verwendet zu werden, und doch genau genug, um eine spürbare Verbesserung der Realitätsnähe zu ermöglichen.

Diese Arbeit ist zwei dafür typischen Fragestellungen gewidmet. Zum einen enthält sie Untersuchungen zur Bestimmung eines dynamischen Verkehrsmodells, das für den Einsatz in der Echtzeitumgebung eines Test-Tools für Mobilfunksysteme geeignet ist. Außerdem beschäftigt sie sich mit der Tarifierung von zukünftigem Internet-Verkehr, insbesondere mit dynamischen Preismodellen, die eine möglichst treffende Berücksichtigung der ständig fluktuierenden aktuellen Marktsituation erlauben und dennoch keine allzugroße Modifikation existierender Protokolle erfordern.

*Zentrale Ergebnisse* konzentrieren sich zum einen auf die Erweiterung eines relativ jungen Verkehrsmodells, die es ermöglicht, vorgegebene statistische Parameter wie auch charakteri-

stische zeitliche Verläufe des zu modellierenden Verkehrs wiederzugeben, und zwar auf einfache und – wie sich herausstellt – sogar automatisch parametrisierbare Weise. Im Bereich der Preismodellierung sind die Neuerungen mehrfacher Natur: Zunächst wird ein Ansatz aus dem Bereich der stochastischen Netze betrachtet, der die Preisverhältnisse zwischen zwei unterschiedlich priorisierten Klassen von Telefongesprächen mit ansonsten identischen Eigenschaften modelliert. Dieses Modell wird für den Fall von Internet-Verkehr mit hohen Bandbreiten und vielen unterschiedlich charakterisierten Klassen verallgemeinert, und zwar durch die Einbeziehung geeigneter asymptotischer Approximationen. Neben einem kurzen Exkurs in die Welt der Verteilten Systeme, insbesondere zur dynamischen Dienstvermittlung in einem CORBA-Trader, beschäftigen sich die verbleibenden Untersuchungen mit der Anpassung eines Auktionsverfahrens, das für die effiziente Vergabe einzelner Ressourcen bekannt ist, auf den realistischeren Fall von Verbindungen, die sich aus vielen Einzelressourcen unterschiedlicher Netzbetreiber zusammensetzen.

Auf diesen eher kursorisch gehaltenen Überblick über die wesentlichen Inhalte der vorliegenden Arbeit folgt nun eine etwas detailliertere Beschreibung der einzelnen Teile. Zu Beginn des ersten Teils führt *Kapitel 2* kurz in die Mobilkommunikation unter GSM sowie die Prinzipien hinter den Testverfahren für GSM-Systeme ein. Dadurch motiviert läßt sich das exemplarisch untersuchte Problem formulieren: die Modellierung einer empirischen Meßreihe zum zeitlichen Verlauf der Zahl an Location Updates über mehrere Tage. Ein summarischer Überblick über etablierte Verkehrsmodellierungsverfahren ergibt, daß sich einzig die Klasse der autoregressiven Prozesse für die Lösung dieser Aufgabe eignen könnte.

Dementsprechend widmet sich *Kapitel 3* erst einmal dieser Modellklasse. Nach einer Einführung in grundsätzliche Aspekte autoregressiver Modellierung und der Darstellung zweier wichtiger Unterklassen, der AR(p)- und der ARMA(p,q)-Prozesse, zeigt der Ansatz von Box-Jenkins den Weg zu effizienter Parametrisierung. Diese wird zunächst “naiv” für die einfachste Klasse, die AR-Prozesse, durchgeführt und bleibt trotz hoher Modellordnung erfolglos. Erst nach dem Einsatz geeigneter Filtermechanismen lassen sich ARMA-Modelle gewinnen, die bei niedriger Ordnung schon deutliche Fortschritte zeigen. Letzten Endes stellt sich aber dennoch heraus, daß die so erzielbaren Resultate unbefriedigend bleiben.

In dieser Situation bietet das in *Kapitel 4* untersuchte TES-Verfahren einen alternativen Lösungsansatz. In seiner zunächst betrachteten Standardform ist es bereits daraufhin konzipiert, gegebene Randverteilungen und Autokorrelationsfunktionen gut nachzubilden. Es zeigt sich, daß die Verallgemeinerung einer in der bisherigen Forschung nur beiläufig behandelten Modellkomponente, der sogenannten “Stitching-Funktion”, zusätzlich auch eine ausgezeichnete Anpassung des zeitlichen Verlaufs der modellierten Kurve an das Original erlaubt. Die bekannt gute Implementierbarkeit des TES-Verfahrens bleibt davon unberührt; zudem stellt sich heraus, daß ein relativ einfacher Ansatz die automatisierte Bestimmung der Modellparameter ermöglicht. Die Anwendung des gefundenen verallgemeinerten Verfahrens auf andersartige Probleme runden die Untersuchung an dieser Stelle ab.

Damit ist das Ende des ersten Teils erreicht – *Kapitel 5* geht noch auf einige weiterführende Fragestellungen ein, bevor sich die Arbeit der dynamischen Preismodellierung im Internet zuwendet. Zur Einführung wirft *Kapitel 6* einen kurzen Blick in die Praxis heutiger Internet-Tarifierung und schlägt ein Schema zur Klassifizierung von Internet-Preismodellen vor. Nach einem Überblick über den Stand der Forschung in diesem Gebiet werden die Anforderungen an ein tragfähiges Preismodell aus Sicht eines Charging-und-Accounting-Tools formuliert. Als wichtiges abstraktes Grundkonzept ergibt sich die Bereitstellung einer "Black Box" an bestimmten Punkten im Netz, in der sich geeignete Preismodelle mit der Bestimmung der aktuellen Marktsituation befassen.

Die folgenden Kapitel versuchen, diesem Konzept auf unterschiedliche Weise Rechnung zu tragen, je nachdem, ob man den Dienstanbieter oder den Kunden als Ausgangspunkt wählt. In ersterem Falle besteht eine denkbare Möglichkeit in der Bestimmung eines Zusammenhangs zwischen der aktuellen Auslastung einer Verbindung und der für die Benutzung zu entrichtenden Gebühr. *Kapitel 7* kehrt hierfür zu mathematischer Grundlagenforschung zurück. Es führt in ein Modell ein, das vor einigen Jahren von Kelly entwickelt wurde, um eine untere Grenze für die Verlustrate in vollvermaschten Telefonnetzen zu bestimmen. Ein seither unbeachtet gebliebener Aspekt dieses Ansatzes erlaubt es, einen Ausdruck für das Preisgefälle zwischen direkt und indirekt gerouteten Telefongesprächen herzuleiten. Die exakte Verallgemeinerung dieses Modells auf den Fall von mehreren Dienstklassen unterschiedlicher QoS-Anforderungen bei hohen verfügbaren Bandbreiten hat sich bislang allerdings als undurchführbar erwiesen. Deshalb greift diese Arbeit zu einer Approximation, die von Mitra et al. eigentlich für das Design von Virtual Private Networks entwickelt und eingesetzt wurde, sich aber als gut geeignet für das vorliegende Problem erweist. Hiermit lassen sich nunmehr zumindest näherungsweise Funktionen bestimmen, die für realistische Kapazitäten wie etwa 622 Mbps und Mehrklassenverkehr den gewünschten Zusammenhang zwischen Auslastung und Preis liefern.

Während der geschilderte Ansatz einzig auf dem momentanen Zustand des Netzes basiert, ist es aber auch denkbar, von der Seite des Nutzers und seiner Präferenzen auszugehen und über entsprechende Auktionsmechanismen den gesuchten Preis zu ermitteln. In *Kapitel 8* wird zunächst allgemein die Darstellung solcher Präferenzen mit Hilfe sogenannter "Utility-Funktionen" diskutiert. Es handelt sich dabei übrigens um ein recht allgemein verwendbares Konzept, das sich beispielsweise auch dafür einsetzen läßt, die Dienstvermittlung im CORBA-Trader eines Verteilten Systems zu dynamisieren, wie in einem Exkurs (Anhang B) genauer demonstriert wird.

Klassische Auktionsverfahren versteigern gewöhnlich ein einzelnes, klar umgrenztes Gut, während es im Fall einer Internet-Verbindung um die Aneinanderreihung mehrerer versteigerter Ressourcen geht. Gehören diese auch noch unterschiedlichen Providern, so sind Synchronisationsprobleme nicht zu vermeiden, zudem kann bereits der Verlust einer einzigen lokalen Auktion zum Zusammenbrechen der kompletten Verbindung führen. Zur Lösung des Problems wird ein neues Schema vorgeschlagen: CHiPS (für: Connection-Holder-is-Preferred-Scheme).

Es erlaubt auf einfache Weise die Wahrung sowohl der Kunden- als auch der Netzbetreiber-Interessen, indem es bestehende Verbindungen für kurze Zeit bevorzugt behandelt. Hierzu sind einige praktische Aspekte zu klären, bevor eine Simulationsstudie das Verhalten des neuen Auktionsmechanismus in Multiprovider-Szenarien genauer untersucht.

*Kapitel 9* widmet sich abschließend noch einigen allgemeineren Ergänzungen zu den beiden vorgestellten Ansätzen und ihrer Anwendung im Rahmen einer Charging-und-Accounting-Plattform, bevor die Schlußbemerkungen von *Kapitel 10* die gesamte Arbeit nochmals kurz zusammenfassen, abrunden und in einen weiterführenden Kontext stellen.

KAPITEL 2

# *Dynamische Verkehrsmodellierung: Einführung und Problemstellung*

## 2.1 Mobilkommunikation unter GSM

Ein erstes zentrales Beispiel für den in dieser Arbeit untersuchten Übergang von statischer zu dynamischer Modellierung, wie er von "praktischer Seite" immer wieder als Wunsch an "die Theorie" herangetragen wird, ist dem Gebiet des Testens von Mobilfunksystemen entnommen. Daher erfolgt in diesem Kapitel zunächst eine kurze Einführung in den exemplarisch betrachteten GSM-Standard. Aus den darauffolgenden Bemerkungen zum Testen von GSM-Systemen erwächst die Notwendigkeit des Einsatzes eines geeigneten dynamischen Verkehrsmodells, wie sie in Abschnitt 2.2 als Referenzproblem spezifiziert wird. Abgerundet wird dieses Kapitel dann durch eine summarische Übersicht von seit längerem etablierten Ansätzen zur Verkehrsmodellierung und deren Einordnung und Bewertung hinsichtlich unserer Problemstellung.

### 2.1.1 Übersicht über den GSM-Standard

Nach etwa einem Jahrzehnt Vorarbeiten wurde 1991 unter der Verantwortung des European Telecommunication Standard Institute ETSI der GSM1800 (Global System for Mobile Communication [MP94]) als europäischer Standard für digitale Telekommunikation herausgegeben; in den Folgejahren bis 1996 wurde er dann zum GSM1900 ausgebaut.

GSM-Netze haben im wesentlichen folgende Struktur (vgl. Abbildung 2-1) [RKJS98]: Das gesamte Versorgungsgebiet besteht aus mehreren MSC-Gebieten, wovon jedes durch ein *Mobile Switching Center (*MSC) verwaltet wird und nochmals in mehrere *Location Areas* (LA) unterteilt ist; jede LA ihrerseits wird über einen *Base Station Controler* (BSC) verwaltet und besteht aus mehreren *Zellen*, in denen jeweils eine *Base Transceiver Station* (BTS) Radiowellen eines bestimmten Frequenzspektrums emittiert und empfängt. Das verteilte Lokationsmanagement besteht aus einer zentralen Datenbank, dem *Home Location Register* (HLR), in dem jeweils das momentane MSC-Gebiet eines Nutzers gespeichert ist, wohingegen für jedes

MSC-Gebiet eine eigene lokale Datenbank, das *Visitor Location Register* (VLR), zuständig ist und die momentane LA eines im überwachten MSC-Gebietes befindlichen Nutzers aufzeichnet.

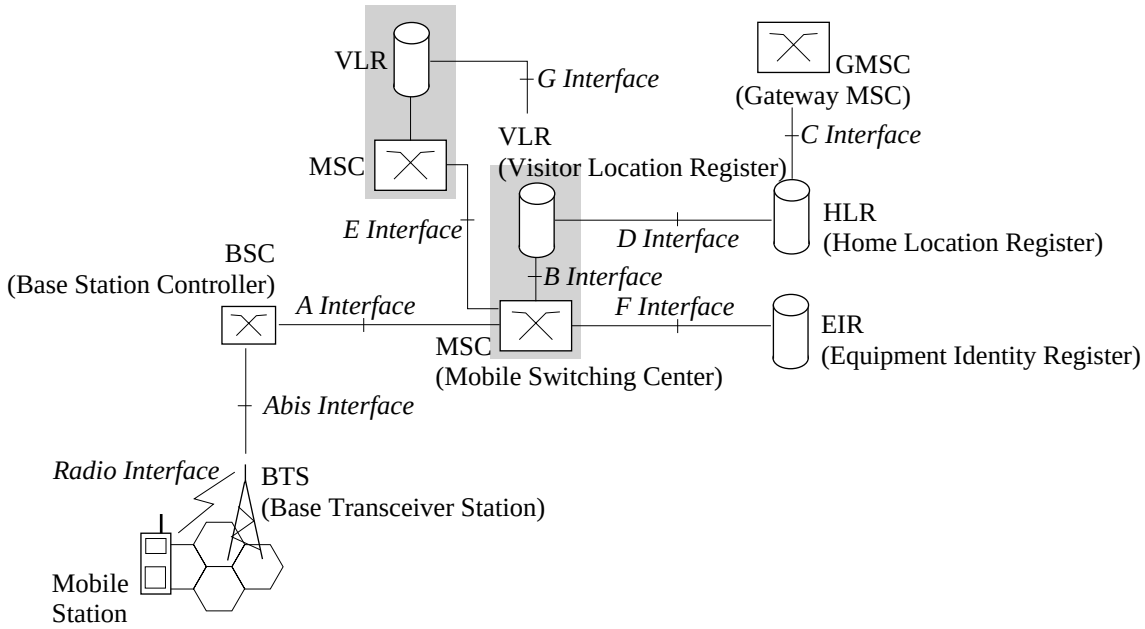


**Abbildung 2-1:** Elemente und Interfaces eines GSM-Netzwerks [HMRS97]

Jeder Nutzer bewegt sich innerhalb des Versorgungsgebietes, wobei das Mobilitätsmanagement hauptsächlich durch Prozeduren für das Aufbau von *Mobile-Originated Calls* (MO) und *Mobile-Terminated Calls* (MT), also Gesprächen von der bzw. zur Mobilstation (MS), sowie für das *Handover* und *Location Update* (LU), also für die Systemänderungen, die ein Ortswechsel des Nutzers von der derzeitigen in eine Nachbarzelle u. U. notwendig macht. Ein Location Update wird dabei von einer Mobilstation ausgelöst, die von einer Location Area in eine andere wechselt; hierbei wird ein Update der Benutzerdaten in den betroffenen Netzdatenbanken (HLR und VLRs) durchgeführt. Solange beide Location Areas zum gleichen MSC-Gebiet gehören, kann dies lokal erfolgen, ansonsten ist ein aufwendigerer Abgleich zwischen dem HLR und beiden betroffenen VLRs notwendig.

Für die Verbindung zum öffentlichen Festnetz ist ein eigenes Mobile Switching Center zuständig, das sogenannte *Gateway MSC* (GMSC). Es erlaubt den Aufbau von Gesprächen aus dem öffentlichen Telefonnetz (*Public Switched Telephone Network* PSTN) an einen GSM-Benutzer. Um ein Gespräch in das richtige MSC-Gebiet zu routen, werden hierzu die entsprechende Informationen aus den Netzdatenbanken (HLR und VLR) abgerufen, das so gefundene zuständige MSC ist dann für den Aufbau der Funkverbindung zwischen Base Transceiver Station und Mobilstation verantwortlich. Eine weitere Datenbank, das *Equipment Identity Register* (EIR), ist schließlich für die Speicherung der Hardware-Informationen der Endgeräte zuständig.

### 2.1.2 GSM-Systemtest

Die Entwicklung von GSM fand vor dem Hintergrund einer rapide ansteigenden Nachfrage nach nomadischem, drahtlosem und mobilem Zugang zu Kommunikationsdiensten statt und hatte die Vereinigung unterschiedlichster Kommunikationswelten zum Ziel. Demzufolge bietet GSM eine Fülle von Diensten für die Sprach- und Datenübertragung, Mehrparteienkommunikation, Intelligente Netzdienste sowie einen *Short Message Service* (SMS).

Dieser Hintergrund hat enorme Auswirkungen auf die Konzeption von Testmethoden für Netzkomponenten wie etwa die Mobile Switching Centers MSC [HMRS97]: Einerseits wächst dadurch die schiere Anzahl von *Test Cases* für die verschiedenen Tests (z.B. *System Tests* und *Type Acceptance Tests*) geradezu dramatisch an, andererseits werden auch die für *Performance Tests* zu berücksichtigende Verkehrsszenarien (*Traffic Mixes)* immer komplexer. Insbesondere wird während eines Performance Tests das System mit einem bestimmten Traffic Mix geladen, der beispielsweise zusammengesetzt sein kann aus Gesprächen von der Mobilstation ins Festnetz, vom Festnetz zur Mobilstation oder von Mobilstation zu Mobilstation (einschließlich der zugehörigen Handover-Prozeduren), Location Updates, Short Messages von der oder zur Mobilstation, Fax- und Daten-Übertragung u.a.m. Die Spezifikation solcher Szenarien zielt darauf ab, sowohl Verkehrszusammensetzungen mit realem Feldcharakter als auch kontrollierte Überlastsituationen im Test zu berücksichtigen. Die Spezifikation selbst erfolgt in einem sogenannten Verkehrsmodell, das im wesentlichen die Anzahl von Operationen beschreibt, die gegen das zu testende System (*System under Test,* SUT) pro Zeiteinheit gerichtet wird. Die entsprechenden Protokolloperationen werden dann von der Testplattform gemäß dem spezifizierten Verkehrsmodell durchgeführt. Abbildung 2-2 skizziert diesen Zusammenhang von zu testendem System SUT und Testplattform.

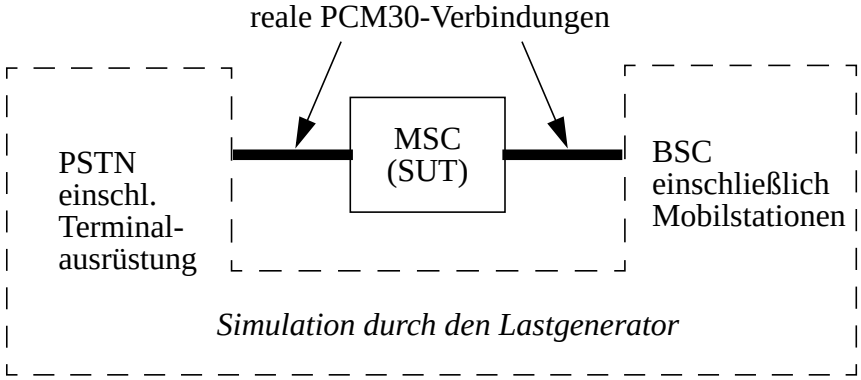


**Abbildung 2-2:** Simulationsumgebung eines Mobile Switching Centers MSC als zu testendem System (System under Test SUT) zwischen öffentlichem Netz (PSTN) und den Basisfunkstationen (BSC)

Hierbei ist die Lasterzeugung innerhalb der Testumgebung in der Regel von einem speziellen Tool, dem *Lastgenerator*, in Echtzeit vorzunehmen. Da nun die erwähnten Traffix Mixes oft hochkomplex sind, verwendete man zur Beschreibung dieser Last auf der realen Verbindung

zum SUT (vgl. Abbildung 2-2) bislang in der Regel statische Verkehrsmodelle [HMRS97]. Offensichtlich ist dies aber weit von jeglicher Realität entfernt, die sich insbesondere durch eine Unzahl statistischer bzw. zeitabhängiger Schwankungen der Last charakterisieren läßt. Ihr kann man sich höchstens nähern, indem man ein wirkliches stochastisches Verkehrsmodell zur Lasterzeugung verwendet, dessen Parametrisierung auf empirisch ermittelten Daten beruht, das diese Messungen in ihrer statistischen Charakteristik weitestgehend reproduziert, dabei allerdings nur minimale Anforderungen an den Modellierungsaufwand stellt und schließlich immer noch in Echtzeit vom Lastgenerator verarbeitet werden kann. Die folgenden Kapitel beschäftigen sich mit der Suche nach einem derartigen Modell. Um diese Untersuchung mehr zu fokussieren, wird allerdings zunächst eine der vielen denkbaren Meßreihen, die für ein Verkehrsszenario zu modellieren sind, herausgegriffen und als “Referenzreihe” eingeführt.

## 2.2 Das Referenzproblem: Dynamische Modellierung von Location Updates im MSC

Abbildung 2-3 zeigt die empirisch ermittelte Anzahl von Location Updates in einem typischen Mobile Switching Center MSC über den Zeitraum von 120 Stunden, in diesem Fall fünf gewöhnlichen aufeinanderfolgenden Werktagen.

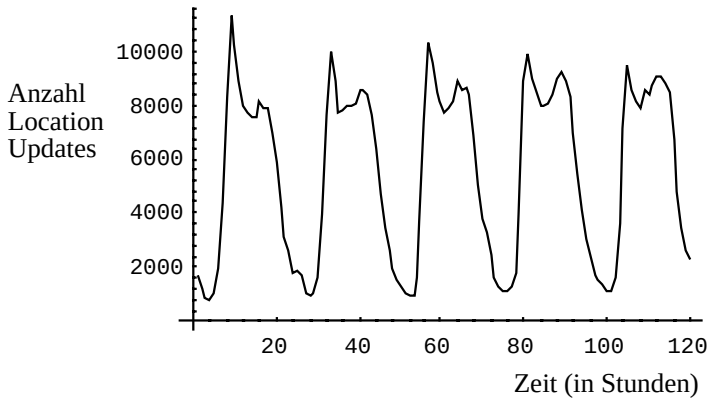


**Abbildung 2-3:** Empirisch gemessener Verlauf von Location Updates über einen Zeitraum von fünf Werktagen [RSH97]

Eine erste Beschreibung dieser Meßreihe fällt nicht sonderlich schwer. Wir beobachten im regelmäßigen Abstand von 24 Stunden eine deutliche Verkehrsspitze, die mit dem Beginn des Arbeitstages zusammenfällt, stellen gegen die Mittagszeit ein typisches Loch fest, von dem sich die Kurve im Laufe des Nachmittags nochmals erholt, bevor Abend und Nacht schließlich für sich auf tiefem Niveau stabilisierende Meßergebnisse sorgen (vgl. hierzu auch [LCW97]).

Mit Mitteln der beschreibenden Statistik ist die Meßreihe noch genauer zu charakterisieren. Hier spielen vor allem zwei Werkzeuge eine wichtige Rolle, und zwar Autokorrelationsfunktion und Randverteilung.

Faßt man die Referenzreihe als einen stationären stochastischen Prozeß[1] $(X_t)_{t = 1, \ldots, N = 120}$ auf, so kann man seinen *Erwartungswert* $\mu_X = E(X_t)$ und damit weiter seine *Autokovarianzfunktion* $\gamma(\tau) = E[(X_t - m_X)(X_{t-\tau} - m_X)]$ zum *"Lag"* (Abstand) $\tau$ bestimmen (und damit die statistische Abhängigkeit zweier Meßpunkte mit zeitlichem Abstand $\tau$ beschreiben). Die *Autokorrelationsfunktion* ergibt sich dann als die mit Lag 0 normierte Autokovarianzfunktion:

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}. \tag{2.1}$$

Aus der Statistik ist bekannt, daß sich der Erwartungswert von $(X_t)_{t = 1, \ldots, 120}$ durch

$$\hat{\mu}_X = \frac{1}{N} \sum_{n=1}^{N} X_n, \tag{2.2}$$

also durch einfache Mittelwertbildung schätzen läßt. Im Falle der Autokovarianzfunktion (und damit auch der Autokorrelationsfunktion) sind unglücklicherweise in der Literatur zwei verschiedene Schätzer verbreitet (vgl. z.B. [SchS89], [Sch93]), die sich allerdings nur in der Normierung auf die Anzahl eingehender Meßwerte unterscheiden:

$$\hat{\gamma}_1(\tau) = \frac{1}{N - \tau} \sum_{n=1}^{N-\tau} X_n X_{n+\tau} \tag{2.3}$$

sowie

$$\hat{\gamma}_2(\tau) = \frac{1}{N} \sum_{n=1}^{N-\tau} X_n X_{n+\tau} \tag{2.4}$$

Hierbei wurde der Einfachheit halber angenommen, daß die Zeitreihe schon auf den Erwartungswert Null transformiert wurde. Beide Schätzer sind konsistent, darüber hinaus ist (2.3) auch erwartungstreu, während sich (2.4) als nur asymptotisch erwartungstreu herausstellt. Im Falle von kleinen Zeitreihen führen diese unterschiedlichen Definitionen typischerweise zu ziemlich stark differierenden Autokorrelationsfunktionen. Abbildung 2-4 links zeigt die beiden Varianten für den Fall der Referenzreihe; dabei ist $\hat{\rho}_1(\tau)$ gestrichelt und $\hat{\rho}_2(\tau)$ durchgehend dargestellt. Wir werden uns im weiteren Verlauf der Arbeit mit beiden Varianten beschäftigen, wobei jeweils klar zu erkennen sein wird, ob es sich um $\hat{\rho}_1(\tau)$ oder $\hat{\rho}_2(\tau)$ handelt.

1. Mathematische Konzepte werden in diesem einführenden Kapitel etwas sorglos verwendet; im späteren Verlauf der Arbeit werden sie - wenn erforderlich - an entsprechender Stelle dann formal korrekt eingeführt. Weiterführende Zusammenhänge sind außerdem noch im Anhang A zusammengefaßt.

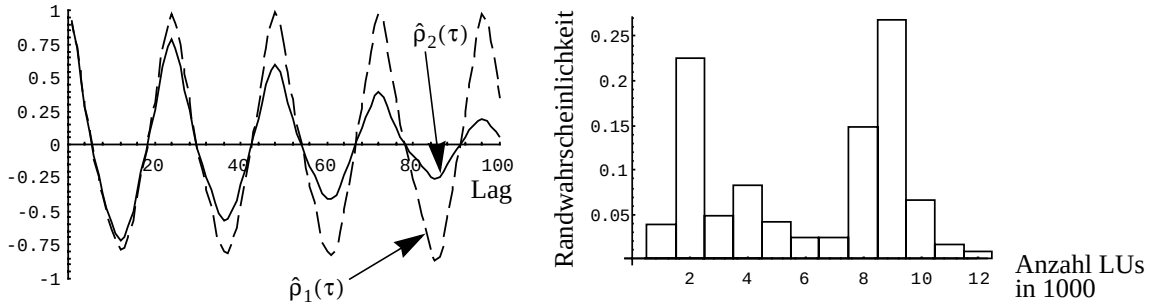


**Abbildung 2-4:** Varianten der empirischen Autokorrelationsfunktion und Randverteilungsdichte (Histogramm) der Referenzmessung

Eine zweite formale Charakterisierung der Referenzreihe bildet ihre *Randverteilung*, deren Dichte in Abbildung 2-4 rechts als Histogramm dargestellt wird. Ihre empirische Ermittlung beruht auf der Einteilung des Wertebereichs der Referenzreihe in diskrete Klassen (hier der Breite 1000) und anschließender Ermittlung der Anzahl von Meßwerten, die in jede Klasse fallen. Wie zu erwarten erhalten wir zwei “Spitzen” (bei 2000 bzw. 9000 LUs), die durch die während der Nacht bzw. während der Arbeitszeit relativ konstant bleibenden Anzahl an LUs zustande kommen.

Nach diesen Vorüberlegungen können wir nun unser Referenzproblem definieren. Es lautet:

***Gesucht ist ein stochastisches Modell für die in Abbildung 2-3 dargestellte Referenzmessung, das ihre statistischen Charakteristiken, insbesondere Autokorrelationsstruktur und Randverteilung, möglichst gut nachbildet und dabei noch die besonderen Voraussetzungen für einen Einsatz im Lastgenerator des beschriebenen GSM-Systemtests erfüllt.***

Zum Abschluß dieses Kapitels werfen wir nun noch einen kurzen Blick auf etablierte Modellierungsverfahren und untersuchen deren Eignung für die eben formulierte Aufgabe.

## 2.3 Wichtige Verfahren der Verkehrsmodellierung

Die Modellierung von Telekommunikationsverkehr durch stochastische Prozesse hat schon eine lange Tradition. Angefangen mit den epochemachenden Untersuchungen des dänischen Mathematikers A. K. Erlang, die 1917 in der Veröffentlichung der nach ihm benannten Formel gipfelten [BHJ48], wurden im Laufe der Zeit eine Fülle von Modellierungsverfahren ausgearbeitet und etabliert. Im folgenden soll ein kurzer Überblick über die wichtigsten Ansätze (insbesondere Erneuerungsprozesse, Markovprozesse, Flußmodelle und Autoregressive Modelle) erfolgen, und zwar anhand der Darstellung im Anhang von [SSRS98]; für eine vertieftere Behandlung vgl. z. B. [FM94], [Kle75] oder [Hav98] bzw. die darin angegebenen Referenzen.

Grundsätzlich erfolgt dabei die Abbildung von Telekommunikationsvorgängen auf stochastische Prozesse durch die Gleichsetzung eines Verbindungsaufbaus mit einer Ankunft im stochastischen System und eines Verbindungsabbaus mit dem Verlassen des Systems. Daraus ergeben sich prinzipiell drei Möglichkeiten mathematischer Beschreibung: die Sequenz $\{T_n\}_{n=1}^{\infty}$ der Ankunftszeiten, die Sequenz der Zwischenankunftszeiten $\{A_n\}_{n=1}^{\infty}$ mit $A_n = T_n - T_{n-1}$ und die Modellierung als Zählprozeß $\{N(t)\}_{t=0}^{\infty}$, wobei $N(t) = max\{n | T_n \leq t\}$ die Anzahl der Ankünfte im Intervall $[0, t]$ ist. Außerdem spielt auch die Verweilzeit im System, d. h. die Zeit zwischen Verbindungsaufbau und Verbindungsabbruch, eine wichtige Rolle bei der Modellierung.

## 2.3.1 Erneuerungsprozesse

Ein Erneuerungsprozeß ist definiert als stochastischer Prozeß, bei dem die Zwischenankunftszeiten $A_n$ i.i.d.[2] mit beliebiger Verteilungsfunktion $F_A(x)$ sind. Der Erneuerungszählprozeß $\{N(t), t \in I\!R\}$ beschreibt die Anzahl der Erneuerungen im Intervall $[0, t]$. In der Regel beschränkt man sich dabei auf die Berechnung seines Erwartungswertes $M(t)$, der sogenannten Erneuerungsfunktion, über die "Fundamental Renewal Equation"

$$M(t) = E[N(t)] = F_A(t) + \int_0^t M(t-s) f_A(s) ds \tag{2.5}$$

### Poissonprozesse

Prominentestes Beispiel für einen Erneuerungsprozeß ist der sogenannte *Poissonprozeß*, der sich durch exponentialverteilte Zwischenankunftszeiten auszeichnet, d.h. $F_A(t) = 1 - e^{-\lambda t}$, $M(t) = \lambda t$ und $m(t) = \frac{dM(t)}{dt} = \lambda$, letzteres die sogenannte Erneuerungsdichte. Die Zeit, bis genau $n$ Erneuerungen stattgefunden haben, ist dadurch Erlang-$n$-verteilt:

$$F_A^{(n)} = 1 - \left( \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} \right) e^{-\lambda t} \tag{2.6}$$

Äquivalent dazu kann der Poissonprozeß als Zählprozeß mit Ankunftsrate $\lambda$ betrachtet werden und weist dann die sogenannte Poissonverteilung

$$P_t(N = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{2.7}$$

auf. Ankünfte in Form eines Poissonprozesses implizieren, daß die Zwischenankunftszeiten exponentialverteilt sind; dies ist unter dem Schlagwort "memoryless-Eigenschaft" bekannt.

2. independent identically distributed

Als gravierende Einschränkung für die sonst mathematisch relativ gut handhabbaren Erneuerungsprozesse ist festzuhalten, daß sie für die Modellierung von Zeitreihen mir signifikanten Autokorrelationen, d.h. Abhängigkeiten innerhalb der Zeitreihe (z.B. von Schritt zu Schritt), aufweisen, aufgrund der Unabhängigkeit der Zwischenankunftszeiten ungeeignet sind.

### 2.3.2 Markovprozesse

Im Gegensatz zu den Erneuerungsprozessen führen Markovprozesse Abhängigkeiten in Zufallssequenzen ein. Dabei ist ein Markovprozeß definiert als stochastischer Prozeß $\{X(t)\}_{t \in \mathbb{N}}$, bei dem die Übergangswahrscheinlichkeit zum nächsten Zustand nur vom aktuellen Zustand des Prozesses abhängt und nicht von weiter zurückliegenden Zuständen:

$$P(X_{t_{n+1}} = x \,|\, X_{t_n} = x_n, X_{t_{n-1}} = x_{n-1}, \ldots, X_{t_0} = x_0) = P(X_{t_{n+1}} = x \,|\, X_{t_n} = x_n). \qquad (2.8)$$

Ist der Wertebereich des Markovprozesses, der sog. Zustandsraum, diskret, so nennt man den Prozeß auch Markov-Kette und unterscheidet weiter Continuous-Time Markov Chains (CTMC) und Discrete-Time Markov Chains (DTMC), je nachdem, ob die Verweilzeiten in den einzelnen Zuständen kontinuierlich oder diskret sind.

Auch hier sei ein bekanntes Exemplar dieser Spezies erwähnt, auf das wir in anderem Zusammenhang in Kapitel 7 noch ausführlich zurückkommen werden: der *Birth-and-Death-Process*, der Übergänge nur zwischen benachbarten Zuständen vorsieht, was die Zurückführung der globalen Betrachtung auf eine Folge von lokalen Berechnungen erlaubt und dadurch die Bestimmung etwa der stationären Zustandswahrscheinlichkeiten sehr erleichtert.

Für eine Modellierung unter besonderer Berücksichtigung der Autokorrelationsfunktion sind einfache Markovprozesse allerdings ebenfalls nicht sonderlich gut geeignet, da ihre Definition nach (2.8) die explizite Abhängigkeit von weiter zurückliegender Vergangenheit ja ausschließt. Etwas bessere Ergebnisse lassen sich erhalten, wenn man einen Markovprozeß gewissermaßen zur Steuerung des eigentlichen Ankunftsprozesses einsetzt, sei es als Markovscher Ankunftsprozeß oder als Markov-modulierter Prozeß.

#### Markovsche Ankunftsprozesse

Der *Markov Arrival Process* (MAP) besteht aus einer CTMC mit endlichem Zustandsraum, welcher aus transienten ($S_T$) und absorbierenden ($S_A$) Zuständen besteht. Übergänge zwischen transienten Zuständen sind durch Raten $C_{ij}$ beschrieben, der Übergang von einem transienten $s_i$ in einen absorbierenden Zustand $s_j$ erfolgt mit Rate $D_{ij}$, wobei gleichzeitig ein Ankunftsereignis ausgelöst und der Prozeß erneut im transienten Zustand $s_j$ gestartet wird.

#### Markov-modulierte Prozesse

Die grundlegende Idee eines *Markov-modulated Process* (MMP) liegt in der expliziten Einführung des Zustandsbegriffs in den Verkehrsstrom. Hierbei wird wiederum eine CTMC $M = \{M(t)\}_{t=0}^{\infty}$ mit Zustandsraum $\{1, 2, \ldots, m\}$ zugrundegelegt. Solange sich die Kette im Zustand $i$ befindet, ist die Ankunftsintensität des eigentlich modellierten Prozesses nur durch diesen Zustand festgelegt, also durch eine Rate $\lambda_i$ und evtl. höhere Momente beschrieben. Wechselt die Markovkette in einen neuen Zustand, so ändert sich auch diese Rate bzw. die höheren Momente entsprechend.

Bekanntester Vertreter dieser Prozeßklasse ist der *Markov-modulated Poisson Process* (MMPP), dessen Zwischenankunftszeiten poissonverteilt mit Rate $\lambda_k$ sind, solange sich die Markovkette im Zustand $k$ befindet. Besteht der Zustandsraum der Kette (auch als *Steuerkette* bezeichnet) nur aus zwei Zuständen $\{0, 1\}$ mit zugehörigen Raten $\lambda_0 = 0, \lambda_1 = \lambda$, spricht man von einem *Interrupted Poisson Process* (IPP) oder auch *On-Off Process*.

### 2.3.3 Flußmodelle

Ein ganz anderer Modellierungsansatz wird bei den Flußmodellen verfolgt: es werden keine Einzelankünfte mehr betrachtet, sondern eine genügend große Anzahl von Ankünften wird zu einem sogenannten *Flow* aggregiert. Dies führt zu einer wesentlichen Senkung der Komplexität, und die numerische Handhabung vereinfacht sich erheblich.

### 2.3.4 Autoregressive Modelle

Autoregressive Modelle wurden in der Zeitreihenanalyse explizit dazu verwendet, Beziehungen innerhalb vorliegender Zeitreihen für die Identifikation und damit letztlich die Modellierung des dahinterliegenden Prozesses auszunützen. Dabei teilen sich die autoregressiven Modelle in mehrere Unterklassen auf (vgl. z.B. [SchS89]):

- Die Klasse der *AR(p)-Prozesse* (*Autoregressive Processes*) besteht aus stochastischen Prozessen der Form

$$X_n = a_0 + \sum_{r=1}^{p} a_r X_{n-r} + U_n \qquad n > 0 \tag{2.9}$$

  mit Zufallsvariablen $X_n$, reellen Gewichten $a_0, \ldots, a_p$ und einem *White-Noise*-Prozeß $U_n$ (d.h. unabhängigen (0,1)-normalverteilten Zufallsvariablen $U_n$).
- Die Klasse der *MA(q)-Prozesse* (*Moving-Average Processes*) hat die Form

$$X_n = \sum_{r=0}^{q} b_r U_{n-r} \qquad n > 0 \tag{2.10}$$

  mit reellen Gewichten $b_0, \ldots, b_q$ und wiederum einem *White-Noise*-Prozeß $U_n$.

- Die Klasse der *ARMA(p,q)-Prozesse* stellt in gewissem Sinn die Kombination der beiden bislang erwähnten dar:

$$X_n = a_0 + \sum_{r=1}^{p} a_r X_{n-r} + \sum_{r=0}^{q} b_r U_{n-r} \tag{2.11}$$

  Ihr Vorteil liegt in der größeren Flexibilität bei wesentlich geringerer Parameterzahl im Vergleich zu AR- und MA-Prozessen.

- Weitere Klassen wie etwa *ARIMA*- und *FARIMA-Prozesse* haben gewisse Bedeutung für die Modellierung von selbstähnlichem Verkehr erlangt, gehen aber über die hier behandelte Zielstellung deutlich hinaus.

Soweit eine kurze Übersicht über die mehr oder weniger “klassischen” Ansätze zur Verkehrsmodellierung. Eine Untersuchung der einzelnen Klassen hinsichtlich der in Abschnitt 2.2 formulierten Problemstellung zeigt relativ schnell, daß – insbesondere im Hinblick auf die explizite Berücksichtigung der Autokorrelationsfunktion – die zuletzt angerissenen autoregressiven Modelle als erste Kandidaten für eine Lösung in Frage kommen. Deshalb wird im nun folgenden Kapitel 3 genauer auf diese Modellklasse eingegangen und dargestellt, welche Resultate sich bei der Modellierung der Referenzreihe aus Abbildung 2-3 durch autoregressive Prozesse erzielen lassen.

KAPITEL 3

# Klassische Autoregressive Zeitreihenmodellierung

## 3.1 Grundlagen autoregressiver Modelle

Bevor wir in den Abschnitten 3.2 und 3.3 im Detail untersuchen, inwieweit sich der in Kapitel 2 skizzierten Problemstellung mit "klassischen" Lehrbuchmethoden beikommen läßt, erfolgt zunächst in aller Kürze deren mathematische Fundierung. Der Schwerpunkt liegt dabei auf der Einführung autoregressiver Prozesse, insbesondere der sogenannten AR- und ARMA-Modelle. Für eine detailliertere Einführung sei z.B. auf [CL66] oder [SchS89] verwiesen.

### 3.1.1 Stationarität und Filter

Damit eine Zeitreihe mit Hilfe autoregressiver Prozesse modelliert werden kann, muß sie in aller Regel gewisse Stationaritätseigenschaften aufweisen, wie in Definition 3.1 beschrieben:

**Definition 3.1:** *Stationarität*
Ein stochastischer Prozeß $(X_t)_{t \in T}$ heißt

- *mittelwertstationär,* falls $\mu_t = \mu$ konstant ist $\forall t \in T$
- *varianzstationär,* falls $\sigma_t^2 = \sigma^2$ konstant ist $\forall t \in T$
- *kovarianzstationär,* falls $\gamma(t_1 + s, t_2 + s) = \gamma(t_1, t_2)$ konstant ist $\forall s, t_1, t_2 \in T$
- *schwach stationär,* falls er mittelwert- und kovarianzstationär ist.

Hierbei ist die *Kovarianz* einer zweidimensionalen Zufallsvariablen $(X_1, X_2)$ wie üblich definiert als $\gamma(X_1, X_2) = E[(X_1 - EX_1) \cdot (X_2 - EX_2)]$.

Ist die untersuchte Zeitreihe nicht schwach stationär, so sind vor der Modellierung die für die Instationarität verantwortlichen Irregularitäten mit Hilfe von linearen Filtern zu beseitigen.

**Definition 3.2:** *Linearer Filter*
Eine lineare Transformation $L$ eines stochastischen Prozesses $(X_t)_{t \in T}$ in einen stochastischen Prozeß $(Y_t)_{t \in T}$ gemäß

$$Y_t = \sum_{k=-\infty}^{\infty} c_k X_{t-k} \tag{3.1}$$

wird als *Linearer Filter* bezeichnet. Der Filter wird durch die Folge seiner Gewichte $(c_k)$ dargestellt. Ist diese Folge absolut summierbar, so heißt auch der Filter *absolut summierbar.*

Ein sehr einfaches Beispiel für einen linearen Filter ist der sogenannte Backshift-Operator:

**Definition 3.3:** *Backshift-Operator*
Der *Backshift-Operator* $B$ ist definiert als der lineare Filter mit der Eigenschaft

$$BX_t = X_{t-1} \tag{3.2}$$

Mit Hilfe des Backshift-Operators läßt sich eine instationären Zeitreihe so filtern, daß sie in eine schwach stationäre übergeführt werden kann. Hierzu betrachten wir zunächst[1] die zum linearen Filter $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ gehörige *charakteristische Gleichung*

$$1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p = 0, \tag{3.3}$$

Setzt man für die komplexen Lösungen $B_i$ voraus, daß $|B_i| \geq 1, \quad i = 1, \ldots, p$, gilt, dann läßt sich der Filter $\Phi(B)$ darstellen als Produkt

$$\Phi(B) = \alpha(B)\kappa(B) \tag{3.4}$$

eines *stabilen Filters* $\alpha(B)$ (d.h. alle Nullstellen liegen außerhalb des Einheitskreises) und eines *grenzstabilen Filters* $\kappa(B)$ (d.h. alle Nullstellen liegen auf dem Einheitskreis).

Dabei lassen sich alle grenzstabilen Filter aus einigen wenigen Bausteinen zusammenfügen:

**Satz 3.4:** *Grenzstabile Filter*
Alle reellen grenzstabilen Filter können als Produkte der Filter $(1 - B)$, $(1 + B)$ und $(1 - 2\cos(2\pi\lambda B) + B^2)$ dargestellt werden.

Durch Anwendung eines geeigneten grenzstabilen Filters auf eine instationäre Zeitreihe läßt sich eine für die Modellierung hinreichende Stationarität erzwingen [Tra98]. Dabei liefert das Verhalten der empirischen Autokorrelationsfunktion zumindest Anhaltspunkte für eine geeignete Wahl von $\kappa(B)$, wie in Tabelle 3-1 zusammengefaßt.Die Anwendung dieser Filter wird in Abschnitt 3.3.1 am praktischen Beispiel demonstriert.

---

1. mit $B^2 X_t = B(B(X_t))$ etc. als der üblichen Konkatenation von Operatoren

| Struktur der empirischen Autokorrelation | anzuwendender Filter |
|---|---|
| (1) langsames Absinken beginnend bei +1 | $(1 - B)$ |
| (2) alternierende Werte nahe ±1 | $(1 + B)$ |
| (3) Peaks bei den Lags $ks$, $k = 1, 2, \ldots$ | $(1 - B^s)$ |
| (4) Schwingungen der Periode $\lambda$ | $(1 - 2\cos(2\pi\lambda B) + B^2)$ |

**Tabelle 3-1:** Zusammenhang zwischen empirischer Autokorrelationsfunktion und grenzstabilen Filtern

Abschließend sei darauf hingewiesen, daß lineare Filter auch noch anders eingesetzt werden können, und zwar zur Darstellung stochastischer Prozesse:

**Satz 3.5:** *Darstellung eines stochastischen Prozesses mit Hilfe eines Filters*

Sei $(X_t)_{t \in T}$ ein stochastischer Prozeß mit $E(X_t^2) \le K < \infty \quad \forall t \in T$ und $(c_k)_{k \in T}$ ein absolut summierbarer Filter. Dann existiert ein stochastischer Prozeß $(Y_t)_{t \in T}$, so daß $\forall t \in T$ gilt:

$$E\left(\left| Y_t - \sum_{k=-n}^{n} c_k X_{t-k} \right|^2\right) \to 0 \quad \text{für} \quad n \to \infty . \tag{3.5}$$

Hiermit läßt sich ein stochastischer Prozeß durch seine lineare Beziehung zu einem White-Noise-Prozeß definieren. Hierzu benötigen wir zunächst

**Definition 3.6:** *White-Noise-Prozeß ("weißes Rauschen")*

Als *White-Noise-Prozeß* oder reinen Zufallsprozeß bezeichnet man allgemein eine Folge $(U_t)_{t \in T}$ von identisch verteilten und unabhängigen Zufallsvariablen $U_t$.

Wir beschränken uns dabei im folgenden auf (0,1)-normalverteilte Zufallsvariable (d.h. solche mit $\mu = 0$ und $\sigma^2 = 1$). Damit ergibt sich

**Definition 3.7:** *Allgemeiner linearer Prozeß*

Seien $(U_t)_{t \in T}$ ein White-Noise-Prozeß und $(c_k)_{k \in T}$ ein absolut summierbarer Filter. Ein stochastischer Prozeß $(X_t)_{t \in T}$ mit

$$X_t = \sum_{k=-\infty}^{\infty} c_k U_{t-k} \tag{3.6}$$

heißt *allgemeiner linearer Prozeß*.

### 3.1.2 AR(p)-Prozesse

**Definition 3.8:** *AR(p)-Prozeß*

Ein stochastischer Prozeß $(X_t)_{t \in T}$ heißt *autoregressiver Prozeß p-ter Ordnung* – kurz: AR(p)-Prozeß – wenn gilt:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + U_t \tag{3.7}$$

wobei $(U_t)_{t \in T}$ ein White-Noise-Prozeß (vgl. Definition 3.6) und $\phi_1, \phi_2, \ldots, \phi_p$ reelle Zahlen (Gewichte) mit $\phi_p \neq 0$ sind.

Der AR(p)-Prozeß $(X_t)_{t \in T}$ heißt *nicht-vorgreifend*, wenn für $k = 1, 2, \ldots$ gilt:

$$Cov(U_t, X_{t-k}) = 0\,. \tag{3.8}$$

Der AR(p)-Prozeß "erinnert" sich also an $p$ vorherige Werte. (3.7) kann man als Regression von $X_t$ bezüglich der $X_{t-i}$ $(i = 1, 2, \ldots, p)$ auffassen. Da die Zufallsvariable $X_t$ von den vorhergehenden Zufallsvariablen $X_{t-i}$ desselben stochastischen Prozesses abhängt, bezeichnet man diese 1927 von Yule eingeführten Prozesse als autoregressiv [Yul27].

Zur Umformung von (3.7) wird über den Backshift-Operator $B$ ein Filter $\Phi(B)$ definiert:

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p\,. \tag{3.9}$$

Damit läßt sich (3.7) schreiben als

$$\Phi(B) X_t = U_t \tag{3.10}$$

Die Anwendung des Filters bewirkt also die Transformation der Zufallsvariablen $X_t$ in ein weißes Rauschen. Auflösung nach $X_t$ ergibt

$$X_t = \Phi^{-1}(B) U_t\,. \tag{3.11}$$

Die Stationarität eines AR(p)-Prozesses ist durch die im folgenden Satz eingeführte *Stationaritätsbedingung* (3.12) festgelegt (für eine Kurzfassung des Beweises vgl. [Tra98]):

**Satz 3.9:** *Stationarität eines AR(p)-Prozesses*

Gegeben sei ein AR(p)-Prozeß $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + U_t$, der die Stationaritätsbedingung erfüllt, d.h. alle Lösungen der charakteristischen Gleichung

$$1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0 \tag{3.12}$$

liegen außerhalb des Einheitskreises. Dann gilt:

1. Der AR(p)-Prozeß $(X_t)_{t \in T}$ läßt sich als linearer Prozeß

$$X_t = \left(1 - \sum_{k=1}^{\infty} b_k B^k\right) U_t = U_t - \sum_{k=1}^{\infty} b_k U_{t-k} \tag{3.13}$$

   mit absolut summierbarer Koeffizientenfolge $(b_k)$ darstellen.

2. Der AR(p)-Prozeß $(X_t)_{t \in T}$ ist schwach stationär und nicht vorgreifend.

Als nächstes werden Erwartungswert und Varianz eines autoregressiven Prozesses $(X_t)_{t \in T}$, der die Stationaritätsbedingung erfüllt, untersucht. Der Erwartungswert lautet mit (3.13)

$$E(X_t) = E(U_t) - \sum_{i=1}^{\infty} b_i E(U_{t-i}) = 0. \tag{3.14}$$

Der Erwartungswert verschwindet somit, d.h. für die Varianz $\sigma^2_{X_t} = \sigma^2_X = E(X_t - E(X_t))^2$ erhält man mit (3.7) (wegen der Stationarität kann der Index $t$ einfach unterdrückt werden)

$$\begin{aligned} \sigma^2_X &= E(X_t^2) = E[(\phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + U_t) \cdot X_t] \\ &= \phi_1 \cdot E(X_{t-1} \cdot X_t) + \ldots + \phi_p \cdot E(X_{t-p} \cdot X_t) + E(U_t \cdot X_t). \end{aligned} \tag{3.15}$$

Weil $(X_t)_{t \in T}$ nicht vorgreifend ist, bleibt nach nochmaligem Einsetzen von (3.7) für den letzten Summanden lediglich der Ausdruck $\sigma^2_U$ übrig, die gesamte Varianz ergibt sich damit als gewichtete Summe der Autokovarianzen $\gamma_X(k)$ $(k = 1, 2, \ldots, p)$ der Zufallsvariablen des AR(p)-Prozesses plus die Varianz des zugehörigen White-Noise-Prozesses:

$$\sigma^2_X = \phi_1 \gamma_X(1) + \phi_2 \gamma_X(2) + \ldots + \phi_p \gamma_X(p) + \sigma^2_U. \tag{3.16}$$

Daher sind als nächstes die Autokovarianzen *k*-ter Ordnung zwischen den Zufallsvariablen $X_t$ und $X_{t-k}$ des AR(p)-Prozesses herzuleiten, mit (3.14) und (3.7) ergibt sich dafür:

$$\begin{aligned} \gamma_X(k) &= E[(X_t - E(X_t))(X_{t-k} - E(X_{t-k}))] = E(X_t \cdot X_{t-k}) - E(X_t) \cdot E(X_{t-k}) \\ &= \phi_1 \cdot E[X_{t-1} X_{t-k}] + \ldots + \phi_p \cdot E[X_{t-p} X_{t-k}] + E(U_t \cdot X_{t-k}). \end{aligned} \tag{3.17}$$

Wiederum ist bei nicht vorgreifenden Prozessen $E(U_t \cdot X_{t-k}) = 0$ für $k > 0$, folglich ist

$$\gamma_X(k) = \phi_1 \cdot \gamma_X(k-1) + \ldots + \phi_p \cdot \gamma_X(k-p). \tag{3.18}$$

Dividiert man (3.18) durch die Varianz $\sigma^2_X$ der Zufallsvariablen $X_t$, so erhält man hieraus die Autokorrelationen *k*-ter Ordnung zwischen den Zufallsvariablen $X_t$ und $X_{t-k}$:

$$\rho_X(k) = \phi_1 \cdot \rho_X(k-1) + \phi_2 \cdot \rho_X(k-2) + \ldots + \phi_p \cdot \rho_X(k-p). \tag{3.19}$$

Mit

$$\rho_X(0) = \frac{\gamma_X(0)}{\sigma_X^2} = \frac{\sigma_X^2}{\sigma_X^2} = 1 \tag{3.20}$$

ergeben sich aus (3.19) für $k = 1, 2, 3, \ldots$ schließlich die sog. *Yule-Walker-Gleichungen*:

$$\begin{aligned}
\rho_X(1) &= \phi_1 + \phi_2 \cdot \rho_X(1) + \phi_3 \cdot \rho_X(2) + \ldots + \phi_p \cdot \rho_X(p-1) \\
\rho_X(2) &= \phi_1 \cdot \rho_X(1) + \phi_2 + \phi_3 \cdot \rho_X(1) + \ldots + \phi_p \cdot \rho_X(p-2) \\
&\ldots \\
\rho_X(p) &= \phi_1 \cdot \rho_X(p-1) + \phi_2 \cdot \rho_X(p-2) + \phi_3 \cdot \rho_X(p-3) + \ldots + \phi_p \\
\rho_X(p+1) &= \phi_1 \cdot \rho_X(p) + \phi_2 \cdot \rho_X(p-1) + \phi_3 \cdot \rho_X(p-2) + \ldots + \phi_p \cdot \rho_X(1) \\
&\ldots
\end{aligned} \tag{3.21}$$

Hierbei wurde ausgenutzt, daß die Autokorrelationen eines schwach stationären Prozesses symmetrisch sind, d.h. $\rho_X(-i) = \rho_X(i)$ für alle $i = 0, 1, 2, \ldots$. Aus den ersten $p-1$ Gleichungen kann man die Autokorrelationen $\rho_X(k)$ für $k = 1, 2, \ldots, p-1$ bestimmen, sodann lassen sich die $\rho_X(k)$ für $k \geq p$ rekursiv gemäß (3.21) berechnen. Dabei ist die Stationaritätsbedingung aus Satz 3.9 hinreichend für die Regularität der Koeffizientenmatrix und damit für die Lösbarkeit des Gleichungssystems.

### 3.1.3 ARMA(p,q)-Prozesse

**Definition 3.10:** *ARMA(p,q)-Prozeß*

Ein stochastischer Prozeß $(X_t)_{t \in T}$ heißt *autoregressiver moving-average-Prozeß der Ordnung (p,q)* – kurz: ARMA(p,q)-Prozeß –, wenn gilt:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + U_t - \theta_1 U_{t-1} - \theta_2 U_{t-2} - \ldots - \theta_q U_{t-q} \tag{3.22}$$

für reelle $\phi_i, \theta_j$ mit $\phi_p \neq 0$, $\theta_q \neq 0$ und $\sum_{k=1}^{p} (\phi_k - \theta_k)^2 > 0$ falls $p = q$.

Wie in Definition 3.8 sei $(U_t)_{t \in T}$ ein White-Noise-Prozeß, zudem wird wie dort $(X_t)_{t \in T}$ als *nicht vorgreifend* bezeichnet, falls $Cov(U_t, X_{t-k}) = 0 \ \forall k \geq 1$ gilt.

Für die Umformung von (3.22) definieren wir analog zu (3.9) einen Filter $\Theta(B)$ als

$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q. \tag{3.23}$$

Damit läßt sich (3.22) schreiben als

$$\Phi(B)X_t = \Theta(B)U_t \tag{3.24}$$

bzw.

$$X_t = \Phi^{-1}(B)\Theta(B)U_t. \tag{3.25}$$

Faßt man $\Phi(B)$ und $\Theta(B)$ als Polynome in einer komplexen Variablen $B$ auf, so soll dabei $\Phi(B) \neq \Theta(B)$ gelten. Analog zu Satz 3.9 ist die Stationaritätsbedingungen (bzw. Invertierbarkeitsbedingung) für den ARMA(p,q)-Prozeß erfüllt, wenn die Nullstellen dieser Polynome $\Phi(B)$ und $\Theta(B)$ (für komplexes $B$) alle außerhalb des Einheitskreises liegen.

Analog zu Satz 3.9 gilt für ARMA-Modelle folgender

**Satz 3.11:**

Gegeben sei ein ARMA(p,q)-Modell, das die Stationaritätsbedingungen erfüllt. Dann läßt sich $(X_t)_{t \in T}$ darstellen als linearer Prozeß

$$X_t = \left(1 - \sum_{k=1}^{\infty} a_k B^k\right) U_t = U_t - \sum_{k=1}^{\infty} a_k U_{t-k}, \tag{3.26}$$

wobei die reellen Koeffizienten $a_k$ der Bedingung $\sum_{k=1}^{\infty} a_k^2 < \infty$ genügen.

Parallel zum Vorgehen in Abschnitt 3.1.2 werden nun Erwartungswert, Varianz und die Autokovarianzen des schwach stationären ARMA(p,q)-Prozesses $(X_t)_{t \in T}$ berechnet.

Man erhält mit (3.26) den Erwartungswert

$$E(X_t) = E\left(U_t - \sum_{k=1}^{\infty} a_k U_{t-k}\right) = 0 \tag{3.27}$$

und die Varianz

$$\gamma_X(0) = \sigma_X^2 = E(X_t^2) - (E(X_t))^2 = E(X_t^2) = 1 + \sum_{k=1}^{\infty} a_k^2, \tag{3.28}$$

wobei sich die $a_k$ durch Einsetzen von (3.13) – dem expliziten Ausdruck für $\Phi^{-1}(B)$ gem. (3.11) – und (3.23) (als dem entsprechenden Ausdruck für $\Theta(B)$) in (3.25), anschließendes Ausmultiplizieren und Koeffizientenvergleich als Funktionen der ARMA-Modellparameter $\phi_i$ und $\theta_j$ bestimmen lassen (Details vgl. [Tra98]).

Die Autokovarianzen lassen sich schließlich analog zu (3.17) folgendermaßen darstellen:

$$\begin{aligned}\gamma_X(k) &= E[(X_t - E(X_t))(X_{t-k} - E(X_{t-k}))] = E(X_t \cdot X_{t-k}) \\ &= \phi_1 E(X_{t-1}X_{t-k}) + \ldots + \phi_p E(X_{t-p}X_{t-k}) + E(U_t \cdot X_{t-k}) \\ &\quad -\theta_1 E(U_{t-1}X_{t-k}) - \ldots - \theta_q E(U_{t-q}X_{t-k}) \\ &= \phi_1\gamma_X(k-1) + \ldots + \phi_p\gamma_X(k-p) + \gamma_{UX}(k) - \theta_1\gamma_{UX}(k-1) - \ldots - \theta_q\gamma_{UX}(k-q).\end{aligned} \tag{3.29}$$

Hier gehen die sogenannten Kreuzkovarianzen $Cov(U_{t-l}, X_{t-k})$ für beliebige ganze Zahlen $t, k, l$ ein, die für den betrachteten schwach stationären Prozeß nur von der Differenz $l - k$ und nicht von der Zeit $t$ abhängig sind und sich deshalb als

$$\gamma_{UX}(k) = Cov(U_t, X_{t-k}) \tag{3.30}$$

darstellen lassen. Nach einer geeigneten Fallunterscheidung [Tra98] erhält man dafür

$$\gamma_{UX}(k) = \begin{cases} 0 & \text{für } k > 0 \\ \sigma_U^2 & \text{für } k = 0 \\ -a_{|k|}\sigma_U^2 & \text{für } k < 0 \end{cases} \tag{3.31}$$

d.h. die auftretenden Kreuzkovarianzen sind entweder gleich Null oder lassen sich als Funktionen in $\sigma_U^2$ und via $a_k$ in den Parametern $\phi_i$ und $\theta_j$ darstellen. Es ist zu beachten, daß im Gegensatz zu den Autokovarianzfunktionen $\gamma_X(k)$ des ARMA-Prozesses die Kreuzkovarianzen $\gamma_{UX}(k)$ nicht mehr symmetrisch in $k$ sind.

Aus (3.29) erhält man nun für $k = 0, 1, 2, \ldots, p$ folgendes Gleichungssystem:

$$\begin{aligned}\gamma_X(0) &= \phi_1\gamma_X(1) + \ldots + \phi_p\gamma_X(p) + \gamma_{UX}(0) - \theta_1\gamma_{UX}(-1) - \ldots - \theta_q\gamma_{UX}(-q) \\ \gamma_X(1) &= \phi_1\gamma_X(0) + \ldots + \phi_p\gamma_X(p-1) + \gamma_{UX}(1) - \theta_1\gamma_{UX}(0) - \ldots - \theta_q\gamma_{UX}(1-q) \\ \gamma_X(2) &= \phi_1\gamma_X(1) + \ldots + \phi_p\gamma_X(p-2) + \gamma_{UX}(2) - \theta_1\gamma_{UX}(1) - \ldots - \theta_q\gamma_{UX}(2-q) \\ &\ldots \\ \gamma_X(p) &= \phi_1\gamma_X(p-1) + \ldots + \phi_p\gamma_X(0) + \gamma_{UX}(p) - \theta_1\gamma_{UX}(p-1) - \ldots - \theta_q\gamma_{UX}(p-q).\end{aligned} \tag{3.32}$$

Unter Berücksichtigung von (3.31) hat man also ein Gleichungssystem mit $p + 1$ Gleichungen und ebensovielen Unbekannten $\gamma_X(0), \gamma_X(1), \ldots, \gamma_X(p)$ erhalten.

Ist $q > p$, so erhält man in gleicher Weise auch noch $\gamma_X(p+1), \ldots, \gamma_X(q)$. Für $k > max(p, q)$ verschwinden schließlich mit (3.31) die Kreuzkovarianzen in (3.32), und die $\gamma_X(k)$ lassen sich (vgl. (3.29)) rekursiv gemäß

$$\gamma_X(k) = \phi_1\gamma_X(k-1) + \ldots + \phi_p\gamma_X(k-p) \tag{3.33}$$

bestimmen.

Division durch $\sigma_X^2$ führt analog zu (3.19) auf die *Yule-Walker-Gleichungen* des betrachteten schwach stationären ARMA(p,q)-Prozesses:

$$\rho_X(k) = \phi_1 \cdot \rho_X(k-1) + \ldots + \phi_p \cdot \rho_X(k-p) \text{ für } k > q. \tag{3.34}$$

Es bleibt festzuhalten, daß im Gegensatz hierzu bei einem AR(p)-Prozeß die entsprechenden Yule-Walker-Gleichungen bereits ab $k > 0$ gelten.

### 3.1.4 Ansatz von Box-Jenkins

In den vorhergehenden Abschnitten wurden gewisse grundlegende Eigenschaften autoregressiver Prozesse eingeführt, die es erlauben, ein AR(p)- bzw. ARMA(p,q)-Modell an eine vorgegebene empirische Zeitreihe (wie etwa die Referenzreihe aus Abbildung 2-3) anzupassen und diese damit zu modellieren. Grundlage dieser Anpassung ist der Ansatz von Box-Jenkins [BJ76], der aus drei Phasen besteht, nämlich der Identifikation eines geeigneten Modells, der Schätzung der zugehörigen Modellparameter und einer Diagnose der Modellgüte.

#### A. Modellidentifikation

Im Falle eines AR(p)-Modells ist die Identifikation (d.h. im wesentlichen die Spezifikation der Ordnung $p$) auf elegante Weise über die partielle Autokorrelationsfunktion möglich.

**Definition 3.12:** *Partielle Autokorrelationen*
Sei $(X_t)_{t \in T}$ ein schwach stationärer stochastischer Prozeß. Die *partiellen Autokorrelationen* $\widehat{\rho}_X(n)$ zwischen $X_t$ und $X_{t-n}$ sind definiert als

$$\widehat{\rho}_X(n) = \frac{D_{12}(n)}{\sqrt{D_{11}(n) \cdot D_{22}(n)}} \tag{3.35}$$

mit $D_{11}(n) \cdot D_{22}(n) \neq 0$.

Dabei sind $D_{11}(n), D_{12}(n), D_{22}(n)$ Determinanten der sogenannten *Korrelationsmatrix*

$$Corr_{(X_t, X_{t-n}, X_{t-1}, X_{t-2}, \ldots, X_{t-(n-1)})} = \begin{bmatrix} 1 & \rho(n) & \rho(1) & \ldots & \rho(n-1) \\ \rho(n) & 1 & \rho(n-1) & \ldots & \rho(1) \\ \rho(1) & \rho(n-1) & 1 & \ldots & \rho(n-2) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho(n-1) & \rho(1) & \rho(n-2) & \ldots & 1 \end{bmatrix}, \tag{3.36}$$

die dadurch entstehen, daß man die $i$-te Zeile und die $j$-te Spalte der Korrelationsmatrix streicht und $D_{ij}(n)$ als Determinante der verbleibenden Restmatrix berechnet.

Partielle Autokorrelationen drücken die lineare Korrelation der Zufallsvariablen $X_t$ und $X_{t-n}$ aus, wobei der Einfluß der dazwischenliegenden Zufallsvariablen $X_{t-1}$, $X_{t-2}$, ..., $X_{t-(n-1)}$ herausgerechnet wird (d.h. diese Zufallsvariablen künstlich konstant gehalten werden). Ihre besondere Bedeutung für die Identifikationsprozedur eines AR-Modells beruht auf folgendem

**Satz 3.13:** *Partielle Autokorrelationsfunktion eines AR(p)-Prozesses*

Die partielle Autokorrelationsfunktion eines schwach stationären AR(p)-Prozesses bricht nach dem Lag $p$ ab, d.h. $\rho_X(k) = 0$ für $k > p$.

Zur Spezifizierung von $p$ schlagen [BJ76] deshalb vor, die empirische Autokorrelationsfunktion $r(k)$ und die empirische partielle Autokorrelationsfunktion $\widehat{\rho}_X(k)$ in Abhängigkeit vom Lag $k$ zu berechnen. Theoretisch sollten nach Satz 3.13 die partiellen Autokorrelationen ab dem Lag $p$ verschwinden. In der Praxis freilich sind die empirischen Werte dafür mit Ungenauigkeiten behaftet, deren Standardfehler in der Größenordnung von $2\sqrt{N}$ ($N$ Länge der empirischen Zeitreihe) angenommen werden kann. Daher bestimmt man die Ordnung $p$ als denjenigen größten Wert $k$, für den $\widehat{\rho}_X(k)$ wesentlich außerhalb des Intervalls $\left[\frac{-2}{\sqrt{N}}, \frac{2}{\sqrt{N}}\right]$ liegt.

Für die Identifikation eines ARMA(p,q)-Modells gibt es leider keine vergleichbare Vorgehensweise. Daher bleibt nichts anderes übrig, als der Reihe nach solange ARMA-Modelle für Kombinationen $(p, q)$ mit $p, q = 1, 2, \ldots$ aufzustellen, bis eine hinreichende Modellierungsgüte erreicht wird.

### B. Schätzung der Modellparameter

Nachdem die Ordnung des Modells festgelegt ist, sind die zugehörigen Modellparameter zu schätzen. Wesentliche Grundlage dafür bieten bei AR- wie ARMA-Modellen die jeweiligen Yule-Walker-Gleichungen (3.21) bzw. (3.34). Die am meisten verbreiteten Schätzverfahren werden in Abschnitt 3.2.2 bzw. 3.3.3 detailliert beschrieben.

### C. Modelldiagnose

Hat man eine vorgegebene empirische Reihe $(x_t)_{t = 1, 2, \ldots, N}$ als AR(p)-Prozeß (3.7) bzw. ARMA(p,q)-Prozeß (3.22) $(X_t)_{t \in T}$ modelliert, dann bleibt noch zu überprüfen, ob die vorgegebene Reihe auch tatsächlich als eine typische Realisierung dieses Prozesses aufgefaßt werden kann. Dies geschieht hier basierend auf den Residualkorrelationen.

Betrachten wir hierfür zunächst im Fall von AR(p) die *Residuen* der Zeitreihe $(x_t)_{t = 1, 2, \ldots, N}$

$$\hat{\varepsilon}_t = x_t - \hat{x}_t = x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} \text{ für } t = 1, \ldots, N, \tag{3.37}$$

(wobei $\hat{x}_t = \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p}$ das Ergebnis des AR(p)-Modells (3.7) bezeichnet, wenn man die entsprechenden Werte $(x_{t-i})_{i = 1, 2, \ldots, p}$ der empirisch gemessenen Reihe einsetzt) sowie ihre Autokorrelationen

$$r(k) = \frac{\sum_{t=1}^{N-k} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=1}^{N} \hat{\varepsilon}_t^2}, \tag{3.38}$$

wobei $\hat{\phi}_i$ die mit Hilfe eines der im vorigen Absatz erwähnten Schätzverfahren berechneten Modellparameter darstellen.

Die empirischen Autokorrelationen der Residualreihe entsprechen nun den empirischen Autokorrelationen eines White-Noise-Prozesses $(\varepsilon_t)_{t \in T}$, die nach Definition 3.6 aber theoretisch gleich Null sind. Wenn das Modell zutrifft, sollte daher $r(k) \approx 0$ für $k \geq 1$ gelten. Hieraus läßt sich folgendes Testkriterium ableiten [BP70]:

**Satz 3.14:** *Portmanteau-Test von Box-Pierce*

Die Box-Pierce Portmanteau-Statistik

$$Q := N \sum_{k=1}^{m} r^2(k) \tag{3.39}$$

ist bei Gültigkeit des geschätzten AR(p)- bzw. ARMA(p,q)-Modells asymptotisch $\chi^2$–verteilt mit $m - p$ Freiheitsgraden, wenn $m$ hinreichend groß gewählt wird:

$$Q \sim \chi^2[m-p]. \tag{3.40}$$

Das Modell ist also zu verwerfen, wenn $Q$ – verglichen mit dem oberen $\alpha$-Quantil einer $\chi^2[m-p]$-Verteilung – zu groß wird. In der Praxis wird dabei $m = \sqrt{N}$ gewählt, $N$ entspricht der Länge der vorliegenden Zeitreihe und $p$ der Ordnung des Modells.

Die Diagnose von ARMA(p,q)-Modellen verläuft strikt analog: Die Autokorrelationen der Residuen $\hat{\varepsilon}_t = x_t - \hat{x}_t = x_t - \hat{\phi}_1 x_{t-1} - \ldots - \hat{\phi}_p x_{t-p} + \hat{\theta}_1 \varepsilon_{t-1} + \ldots + \hat{\theta}_q \varepsilon_{t-q}$ werden nach (3.38) berechnet. Das Modell ist wiederum dann zu verwerfen, wenn $Q$ – diesmal verglichen mit dem oberen $\alpha$-Quantil einer $\chi^2[m-(p+q)]$-Verteilung – zu groß wird.

Für eine vertiefte Darstellung dieser Statistik wird auf [SchS89] verwiesen.

## 3.2 Modellierung mit AR-Prozessen

Im folgenden werden die theoretischen Resultate, wie sie in Abschnitt 3.1 zusammengefaßt wurden, auf die konkrete Fragestellung nach einem autoregressiven Modell für die vorgegebene empirische Zeitreihe $(x_t)_{t=1,2,\ldots,120}$ der Abbildung 2-3 angewandt. Zunächst erfolgt

dabei eine Untersuchung von in Frage kommenden AR(p)-Modellen, wobei der Einfachheit halber die Referenzreihe in noch ungefilterter Form betrachtet wird. Abschnitt 3.3 wird dann ein ARMA-Modell für die geeignet gefilterte Referenzreihe als das beste Ergebnis ableiten, das unter den autoregressiven Modellen der untersuchten Zeitreihe gefunden werden konnte.

### 3.2.1 Modellidentifikation

Wie in Abschnitt 3.1.4 dargelegt wurde, besteht der erste Schritt in der Anpassung eines AR(p)-Modells an die Referenzreihe in der Bestimmung der Ordnung $p$. Schon der graphischen Darstellung der Referenzkurve (Abbildung 2-3) und des zugehörigen Korrelogramms (Abbildung 2-4 links) läßt sich eine Periodizität der Länge 24 entnehmen; genauer gesagt bestehen starke positive Korrelationen zwischen Werten, die um 24 verschoben sind, während zwischen um 12 verschobene Werte starke negative Korrelationen festzustellen sind. Diese Periodizität wird durch das entsprechende Periodogramm[2] (Abbildung 3-2) noch bestärkt. Es ist deutlich ein starker Peak bei der Frequenz $\lambda = 1/24 = 0.04167$ zu erkennen. Weitere Peaks sind im Periodogramm auch noch für die Frequenzen $\lambda = 1/12 = 0.08333$ bzw. $\lambda = 1/6 = 0.125$ zu erkennen.

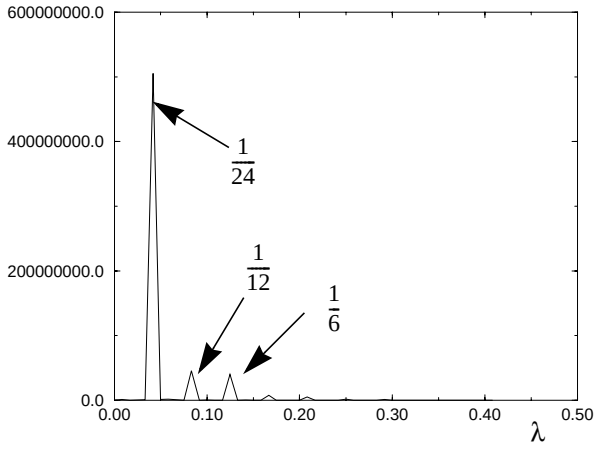


**Abbildung 3-2:** Periodogramm der Referenzreihe

In Abschnitt 3.1.4 wurde bereits dargestellt, wie anhand der partiellen Autokorrelationen ein Kriterium für die Wahl der Ordnung angegeben werden kann. Abbildung 3-3 zeigt die nach Definition 3.12 berechneten partiellen Autokorrelationen der Referenzreihe samt dem zugehörigen Konfidenzintervall $\left[\frac{-2}{\sqrt{120}}, \frac{2}{\sqrt{120}}\right] = [-0.1825, 0.1825]$.

2. Unter dem *Periodogramm* (auch *Stichprobenspektrum* genannt) versteht man eine Funktion $I(\lambda)$ der Frequenz, die für jede Frequenz $\lambda$ angibt, mit welcher Intensität harmonische Wellen dieser Frequenz in der betreffenden Zeitreihe auftauchen. Formal definiert ist es [SchS89] durch

$$I(\lambda) = N\left[\frac{1}{N}\sum_{t=1}^{N}(X_t - \bar{X})\cos 2\pi\lambda t\right]^2 + N\left[\frac{1}{N}\sum_{t=1}^{N}(X_t - \bar{X})\sin(2\pi\lambda t)\right]^2.$$

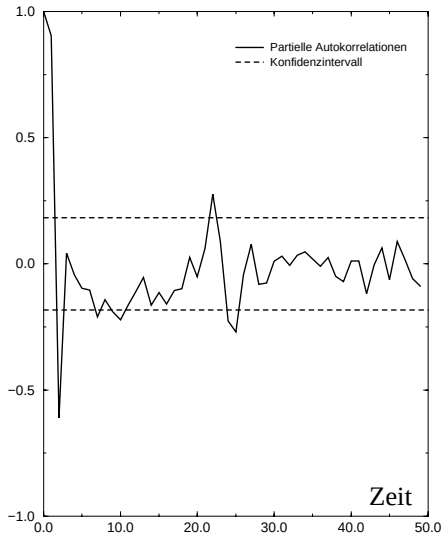


**Abbildung 3-3:** Partielle Autokorrelationen der Referenzreihe und Konfidenzintervall

Mit Hilfe von Satz 3.13 läßt sich nun daraus die Ordnung des AR(p)-Prozesses ablesen. Der größte Wert $k$, für den die partielle Autokorrelation außerhalb des Konfidenzintervalls liegt, ergibt sich demnach zu $k = 24$ und dient im folgenden als Ordnung $p$.

## 3.2.2 Parametrisierungsverfahren

Es sind also nun anhand der $N = 120$ Beobachtungen $x_1, \ldots, x_N$ die $p = 24$ Modellparameter $\phi_1, \phi_2, \ldots, \phi_p$ eines AR(24)-Prozesses $(X_t)_{t \in T}$ gemäß (3.7) zu bestimmen. Hierzu werden im folgenden vier Schätzverfahren angegeben, die allesamt implementiert wurden; in Abschnitt 3.2.3 werden diese dann exemplarisch auf die ungefilterte Referenzreihe angewendet und verglichen. Wir gehen dabei o.B.d.A. von einer mittelwertbereinigten Zeitreihe $(x_t - \bar{x})_{t = 1, \ldots, N}$ aus.

### (a) Yule-Walker-Verfahren

Das erste Verfahren (vgl. [SchS89]) basiert auf den Yule-Walker-Gleichungen (3.21). Die Modellparameter $\phi_1, \phi_2, \ldots, \phi_p$ des AR(p)-Modells sind dabei durch Schätzwerte $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$ derart zu ersetzen, daß eine gute Anpassung des Modells an die gegebenen Zeitreihendaten im Sinne einer Minimierung der Summe der quadrierten Residuen bezüglich $\phi_1, \phi_2, \ldots, \phi_p$ erreicht wird:

$$s^2 = S^2(\phi_1, \ldots, \phi_p) = \sum_{t = -\infty}^{N} \hat{\varepsilon}_t^2 \stackrel{!}{=} \min \tag{3.41}$$

mit den Residuen

$$\hat{\varepsilon}_t = (X_t - \mu) - \phi_1(X_{t-1} - \mu) - \ldots - \phi_p(X_{t-p} - \mu) \tag{3.42}$$

Mit Hilfe der empirischen Autokorrelationen $r_k = r(k)$ wird ein empirisches Analogon zu den Yule-Walker-Gleichungen hergeleitet:

$$\begin{bmatrix} 1 & r_1 & \dots & r_{p-2} & r_{p-1} \\ r_1 & 1 & \dots & r_{p-3} & r_{p-2} \\ \dots & \dots & \dots & \dots & \dots \\ r_{p-1} & r_{p-2} & \dots & r_1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \dots \\ \hat{\phi}_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \dots \\ r_p \end{bmatrix}. \tag{3.43}$$

Die Lösungen $\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p$ werden als *Yule-Walker-Schätzer* bezeichnet. Dieses Verfahren ist sehr einfach in der Anwendung, da die empirischen Autokorrelationen in der Praxis bereits für die Modellidentifikation bestimmt wurden. (3.43) läßt sich mit dem klassischen Gauß-Seidel-Algorithmus [Sto93] oder ähnlichen bekannten numerischen Verfahren lösen.

### (b) Householder-Verfahren

Eine Alternative zu den Yule-Walker-Schätzern sind Kleinste-Quadrate-Schätzer, welche auf der direkten Auswertung des überbestimmten linearen Gleichungssystems in allen $N - p$ übrigen empirischen Autokorrelationen beruhen [SchS89]. Hierbei werden diejenigen Koeffizienten $\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p$ bestimmt, welche die Summe

$$\sum_{t=p+1}^{N} \left( x_t - \sum_{\tau=1}^{p} \hat{\phi}_\tau x_{t-\tau} \right)^2 \stackrel{!}{=} \min \tag{3.44}$$

minimieren. Mit

$$A := \begin{bmatrix} x_p & x_{p-1} & \dots & x_1 \\ x_{p+1} & x_p & \dots & x_2 \\ \dots & \dots & \dots & \dots \\ x_{2p} & x_{2p} & \dots & x_p \\ x_{2p+1} & x_{2p} & \dots & x_{p+1} \\ \dots & \dots & \dots & \dots \\ x_{N-1} & x_{N-2} & \dots & x_{N-p} \end{bmatrix} \tag{3.45}$$

$$c := (\phi_1, \phi_2, \dots, \phi_p) \tag{3.46}$$

$$f := (x_{p+1}, x_{p+2}, \dots, x_N) \tag{3.47}$$

läßt sich (3.44) dann wie folgt schreiben:

$$\| f - Ac \|_2^2 \stackrel{!}{=} \min. \tag{3.48}$$

Zur Lösung des überbestimmten linearen Gleichungssystems $Ac = f$ eignet sich am besten das bekannte Householder-Verfahren [Sto93], da es die Kondition der Matrix $A$ nicht verschlechtert und damit trotz der Überbestimmtheit numerisch stabil bleibt.

### (c) Levinson-Durbin-Rekursion

Bei diesem Ansatz (vgl. [SchS89]) erfolgt die Parameterschätzung auf einfache Weise rekursiv aus den Lösungen des um eine Gleichung reduzierten Yule-Walker-Gleichungssystems mittels des folgenden Algorithmus:

Seien $r_1, \ldots, r_p$ die empirischen Autokorrelationen.

Startwerte: $\hat{\phi}_{1,1} := r_1 \qquad Q(1) := 1 - r_1^2$.

Für $k = 2, \ldots, p$:

(1) $\hat{\phi}_{k,k} = \dfrac{\Delta(k)}{Q(k-1)}$ mit

$$\Delta(k) := r_k - (\hat{\phi}_{1,k-1} r_{k-1} + \hat{\phi}_{2,k-2} r_{k-2} + \ldots + \hat{\phi}_{k-1,k-1} r_1)$$

(2) $\hat{\phi}_{i,k} = \hat{\phi}_{i,k-1} - \hat{\phi}_{k,k} \hat{\phi}_{k-i,k-1}$ für $i = 1, \ldots, k-1$

(3) $Q(k) = Q(k-1) \cdot (1 - \hat{\phi}_{k,k}^2)$

Für die Parameter des AR(p)-Modells erhält man dann die Schätzer

$$\hat{\phi}_i := \hat{\phi}_{ip} \text{ für } i = 1, \ldots, p. \tag{3.49}$$

### (d) Marquardt-Technik für AR-Modelle

Auch das letzte vorgestellte Verfahren basiert auf der Minimierung der Summe der quadratischen Residuen. Aufgrund seiner Komplexität wird es hier nur summarisch und in Anlehnung an [Tra98] vorgestellt; ausführlichere Darstellungen finden sich etwa in [Mar63] oder [BJ76].

1. *Bestimmung der Residuen* $\hat{\varepsilon}_t = x_t - \phi_1^{(i)} x_{t-1} - \ldots - \phi_p^{(i)} x_{t-p}$, $t \in \{1-r, \ldots, N\}$.

   Dabei gibt der obere Index $(i)$ die Iterationsstufe des Verfahrens an. Als Anfangsschätzwerte $\hat{\phi}_1^{(0)}, \ldots, \hat{\phi}_p^{(0)}$ setzt man die Yule-Walker-Schätzer aus Verfahren (a).

   Für die Bestimmung der $\hat{\varepsilon}_t$ für $t = 1-r, \ldots, 1$ benötigt man (nicht-beobachtbare) Startwerte $x_t$ mit $t \le 0$. Diese kann man auf zwei Weisen mit einem Kleinste-Quadrate-Ansatz festlegen: Der *Conditional Least Squares*-Ansatz (CLS) setzt die fraglichen $x_t = \hat{\varepsilon}_t = 0$, $t \le 0$ und beginnt daraufhin die rekursive Bestimmung der übrigen $\hat{\varepsilon}_t$, während der *Unconditional Least Squares*-Ansatz (ULS) die Startwerte mittels einer Backforecast-Technik

[BJ76] iterativ aus den vorhandenen Daten schätzt (für eine detailliertere Darstellung beider Ansätze sei auf [Tra98] oder [HEK98] verwiesen).

2. *Numerische Berechnung der partiellen Ableitung* $\frac{\partial}{\partial \phi_k^{(i)}} \hat{\varepsilon}_t(\phi_1^{(i)}, \phi_2^{(i)}, \ldots, \phi_p^{(i)})$.

   Definiere die $((N+r) \times p)$-Matrix $D = (d_{tk})$ mit sehr kleinem $\delta$ (z.B. $\delta = 0.01$) durch

$$d_{tk}^{(i)} = \frac{\hat{\varepsilon}_t(\phi_1^{(i)}, \ldots, \phi_{k-1}^{(i)}, \phi_k^{(i)} + \delta, \phi_{k+1}^{(i)}, \ldots, \phi_p^{(i)}) - \hat{\varepsilon}_t(\phi_1^{(i)}, \ldots, \phi_{k-1}^{(i)}, \phi_k^{(i)}, \phi_{k+1}^{(i)}, \ldots, \phi_p^{(i)})}{\delta} \quad (3.50)$$

3. *Lösung des überbestimmten linearen Gleichungssystems* $D^{(i)} \cdot x^{(i)} = \hat{\varepsilon}_t(\phi^{(i)})$.

   Hierzu wird statt der schlecht konditionierten Normalgleichungen wiederum die Householder-Transformation [Sto93] verwendet.

4. *Bestimmung der neuen Schätzwerte für die Modellparameter nach der* $(i+1)$*-ten Iteration*:

$$(\hat{\phi}_1^{(i+1)}, \ldots, \hat{\phi}_p^{(i+1)}) = (\hat{\phi}_1^{(i)}, \ldots, \hat{\phi}_p^{(i)}) - (x_1^{(i)}, \ldots, x_p^{(i)}). \quad (3.51)$$

5. *Rücksprung zu Iterationsschritt 1.*

Die Iteration wird abgebrochen, wenn die durch $x^{(i)}$ dargestellte Änderung der Schätzwerte genügend klein geworden ist, also wenn $max\left|x_k^{(i)}\right| < \varepsilon$ für hinreichend kleines $\varepsilon$ gilt. Als Schätzer für die Residualvarianz erhält man dann mit (3.41)

$$\hat{\sigma}_\varepsilon^2 = \frac{S^2(\hat{\phi}_1^{(i)}, \ldots, \hat{\phi}_p^{(i)})}{N-p}. \quad (3.52)$$

### 3.2.3 Beispiel: Ungefilterte Modellierung der Referenzreihe

Die vier in Abschnitt 3.2.2 erläuterten Verfahren wurden implementiert und verglichen. Zur Illustration wenden wir sie in diesem Abschnitt einmal auf die ungefilterte Referenzreihe von Abbildung 2-3 an, was dem naiven Versuch entspricht, durch ein möglichst simples klassisches Verfahren bereits an eine auch in der Praxis zufriedenstellende Modellierung dieser Zeitreihe zu kommen. Die abschließende Anwendung des Diagnoseverfahrens von Satz 3.14 wird eine diesbezügliche Bewertung erlauben.

Die Ordnung des zu bildenden AR-Modells wurde ja bereits in Abschnitt 3.2.1 zu $p = 24$ bestimmt. Die Schätzungen der entsprechenden Modellparameter $\phi_1, \phi_2, \ldots, \phi_{24}$, wie sie sich aus der Anwendung der Verfahren (a) – (d) ergeben, sind in Tabelle 3-4 dargestellt.

Da das Yule-Walker-Verfahren und die Levinson-Durbin-Rekursion beide auf einer direkten Auswertung der Yule-Walker-Gleichungen (3.21) basieren, ist es keine Überraschung, daß sie zu identischen Resultaten für die geschätzten Modellparameter führen, während die beiden

| | Yule-Walker | Householder | Levinson | Marquardt |
|---|---|---|---|---|
| $\hat{\phi}_1$ | 1.201050 | 1.176566 | 1.201050 | 1.180883 |
| $\hat{\phi}_2$ | -0.494221 | -0.533378 | -0.494221 | -0.536751 |
| $\hat{\phi}_3$ | -0.069509 | -0.013485 | -0.069509 | -0.015435 |
| $\hat{\phi}_4$ | 0.031988 | 0.061637 | 0.031988 | 0.070970 |
| $\hat{\phi}_5$ | -0.053626 | -0.042042 | -0.053626 | -0.050240 |
| $\hat{\phi}_6$ | 0.058130 | -0.066770 | 0.058130 | -0.062303 |
| $\hat{\phi}_7$ | -0.100227 | 0.023190 | -0.100227 | 0.024691 |
| $\hat{\phi}_8$ | -0.019618 | -0.065000 | -0.019618 | -0.069745 |
| $\hat{\phi}_9$ | 0.008754 | 0.043513 | 0.008754 | 0.048103 |
| $\hat{\phi}_{10}$ | -0.052701 | -0.082875 | -0.052701 | -0.086600 |
| $\hat{\phi}_{11}$ | -0.013581 | 0.011693 | -0.013581 | 0.017050 |
| $\hat{\phi}_{12}$ | -0.132826 | -0.108061 | -0.132826 | -0.113252 |
| $\hat{\phi}_{13}$ | 0.063605 | 0.052927 | 0.063605 | 0.057308 |
| $\hat{\phi}_{14}$ | -0.095217 | -0.088417 | -0.095217 | -0.088907 |
| $\hat{\phi}_{15}$ | 0.047222 | 0.049492 | 0.047222 | 0.048678 |
| $\hat{\phi}_{16}$ | -0.089770 | -0.080556 | -0.089770 | -0.077373 |
| $\hat{\phi}_{17}$ | 0.020968 | 0.054156 | 0.020968 | 0.051166 |
| $\hat{\phi}_{18}$ | -0.127962 | -0.200425 | -0.127962 | -0.190422 |
| $\hat{\phi}_{19}$ | 0.097208 | 0.187506 | 0.097208 | 0.172765 |
| $\hat{\phi}_{20}$ | 0.044485 | -0.040819 | 0.044485 | -0.038439 |
| $\hat{\phi}_{21}$ | -0.231426 | -0.146286 | -0.231426 | -0.141545 |
| $\hat{\phi}_{22}$ | 0.048982 | -0.026188 | 0.048982 | -0.025245 |
| $\hat{\phi}_{23}$ | 0.356332 | 0.416691 | 0.356332 | 0.415898 |
| $\hat{\phi}_{24}$ | -0.226640 | -0.207800 | -0.226640 | -0.207462 |

**Tabelle 3-4:** Resultate der Schätzverfahren (a) – (d) für die ungefilterte Referenzreihe

anderen Verfahren davon sowie auch untereinander durchaus Abweichungen zu verzeichnen haben (insbesondere sind Vorzeichenwechsel keine Seltenheit).

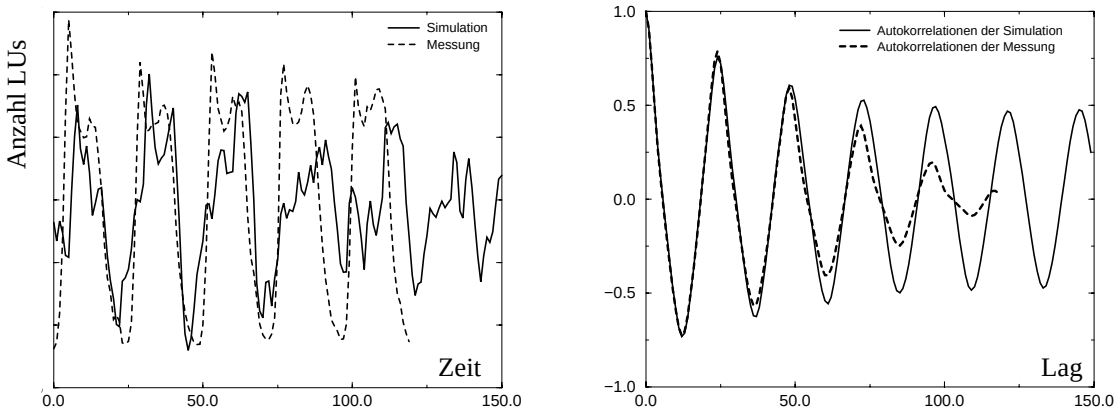


**Abbildung 3-5:** AR(24)-Modell der Referenzreihe nach dem Yule-Walker-Verfahren und zugehörige Autokorrelationsfunktion

Abbildung 3-5 zeigt das AR(24)-Modell der ungefilterten Referenzreihe nach dem Yule-Walker-Verfahren (bzw. nach dem eben Ausgeführten auch dem Levinson-Durbin-Ansatz) und die zugehörige Autokorrelationsfunktion sowie als Vergleich dazu jeweils die entsprechenden Kurven der Referenzreihe. Immerhin wird die Periodizität der Referenzkurve halbwegs gewahrt, und auch die Autokorrelationen niedriger Ordnung stimmen in etwa überein, für höhere Ordnungen bleibt jedoch das Modell deutlich höher autokorreliert.

Dieser Eindruck ändert sich im Fall der beiden übrigen Verfahren nur leicht, wie Abbildung 3-6 zu entnehmen ist. Allerdings ist hier für die ersten Autokorrelationsordnungen ein leichter Unterschied zwischen Modell und Referenzkurve zu bemerken, immerhin stimmen dafür aber die höheren Ordnungen insofern besser überein, als die Amplitude der Autokorrelationsfunktionen der Modelle analog zur Referenzfunktion immer weiter sinkt, was beim Yule-Walker-Verfahren so nicht der Fall ist. Unterschiede zwischen den Kurven für das Householder- und das Marquardtverfahren selbst sind allerdings mit bloßem Auge kaum festzustellen (Abbildung 3-6).

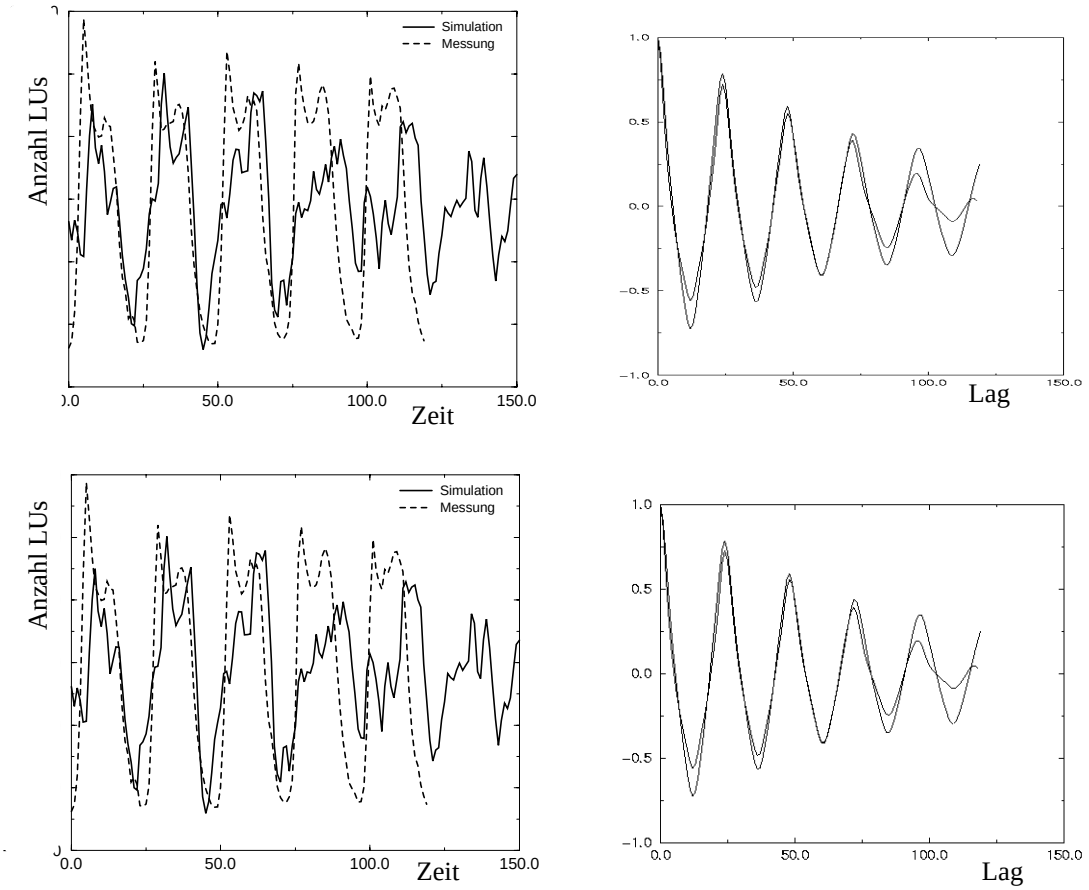


**Abbildung 3-6:** AR(24)-Modelle nach dem Householder- (oben) und Marquardt-Verfahren (unten) und zugehörige Autokorrelationsfunktionen

Diese eher qualitativen Aussagen lassen sich – wie bereits am Schluß von Abschnitt 3.1.4 erläutert – mit Hilfe der Box-Pierce Portmanteau-Statistik auch noch quantifizieren. Hierzu berechnet man zunächst nach (3.37) für jedes der vier Verfahren die Residuen mit den entsprechend geschätzten Parametern, um sodann mit (3.39) den Wert von $Q$ zu bestimmen. Dieser wird schließlich mit dem oberen $\alpha$-Quantil der $\chi^2[m-p]$-Verteilung verglichen. Nach Satz 3.14 ist dabei $m$ hinreichend groß zu wählen, was – wie sich herausgestellt hat – in unserem Fall bedeutet, daß der vorgeschlagene Wert von $m = \sqrt{120} \approx 11$ nicht ausreicht, da die Ordnung $p = 24$ des AR-Verfahrens sehr hoch ist. Deshalb wurde stattdessen $m - p = 11$ gewählt. Das $\alpha$-Quantil einer $\chi^2[11]$-Verteilung ist z.B. in [Sch93] zu finden.

| | **Yule-Walker (a, c)** | **Householder (b)** | **Marquardt (d)** |
|---|---|---|---|
| Q | 8.2978 | 3.755 | 3.8299 |
| $\alpha$-Quantil | 0.3 | 0.025 | 0.025 |

**Tabelle 3-7:** Box-Pierce Portmanteau-Statistik für die Beispielsmodelle

Tabelle 3-7 erlaubt in Verbindung mit Satz 3.14 die Feststellung, daß die Modellgüte nicht befriedigend ist. Dies kann an der Tatsache liegen, daß eine hohe Anzahl von Modellparametern, nämlich 24, anhand einer relativ kleinen Zahl von vorliegenden Meßwerten der Referenzreihe ($N = 120$) zu schätzen waren. Es stellt sich außerdem die Frage, ob das Ergebnis nicht durch Anwendung einer geeigneten Filterung noch zu verbessern wäre. Diesbezüglich durchgeführte neue Anpassungen hatten jedoch keine wesentliche Änderung der Modellgüte zur Folge [Tra98]. Nach dem Scheitern der Anpassung eines AR-Modells wird daher im folgenden Abschnitt nun ein geeignetes ARMA(p,q)-Modell bestimmt, das ein wesentlich besseres Ergebnis der Modellierung zuläßt.

## 3.3 Modellierung mit ARMA-Prozessen

Vor der Bestimmung des eigentlichen ARMA(p,q)-Modells wird zunächst gezeigt, wie sich – aufbauend auf den in Abschnitt 3.1.1 beschriebenen Zusammenhängen – durch den Einsatz geeigneter Filter die Stationaritätseigenschaften der Referenzreihe $(x_t)_{t = 1, 2, \ldots, 120}$ von Abbildung 2-3 verbessert werden können. Nach der Modellidentifikation werden dann zwei Verfahren zur Parameterschätzung sowie die damit erzielbaren Resultate vorgestellt.

### 3.3.1 Filterung

Für die Zuordnung des passenden Filters mit Hilfe von Tabelle 3-1 ist die Struktur der Autokorrelationsfunktion entscheidend. Die bereits in Abschnitt 3.2.1 mit Hilfe des Periodogramms beschriebene Periodizität der Frequenz $\lambda = 1/24$ deutet dabei auf einen Filter vom Typ (4) (in Tabelle 3-1) hin.

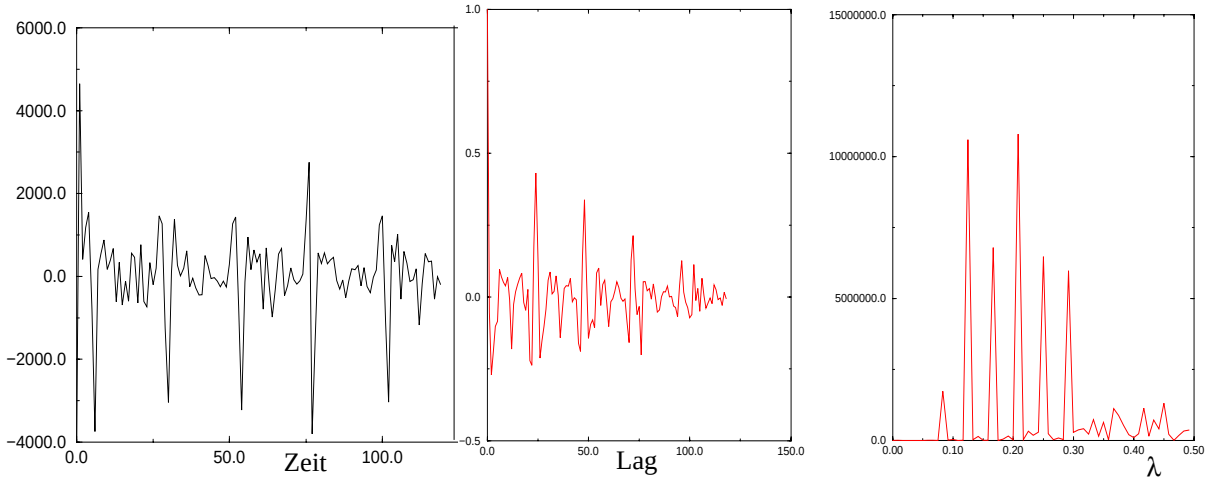


**Abbildung 3-8:** Erste Filterung der Referenzreihe, Autokorrelationen und Periodogramm

Die entsprechend gefilterte Zeitreihe

$$y_t = \left(1 - 2\cos\left(\left(2\pi\frac{1}{24}\right)B + B^2\right)\right) x_t \tag{3.53}$$

mit Backshift-Operator $B$ und $t = 1, \ldots, 120$ ist samt zugehöriger Autokorrelationsfunktion und Periodogramm in Abbildung 3-8 dargestellt. Die Autokorrelationsfunktion weist nun deutliche Peaks im Abstand 24 auf, das Periodogramm zeigt keine Auffälligkeiten. Deshalb erfolgt nun noch die Anwendung eines Filters vom Typ (3):

$$z_t = (1 - B^{24})\, y_t \tag{3.54}$$

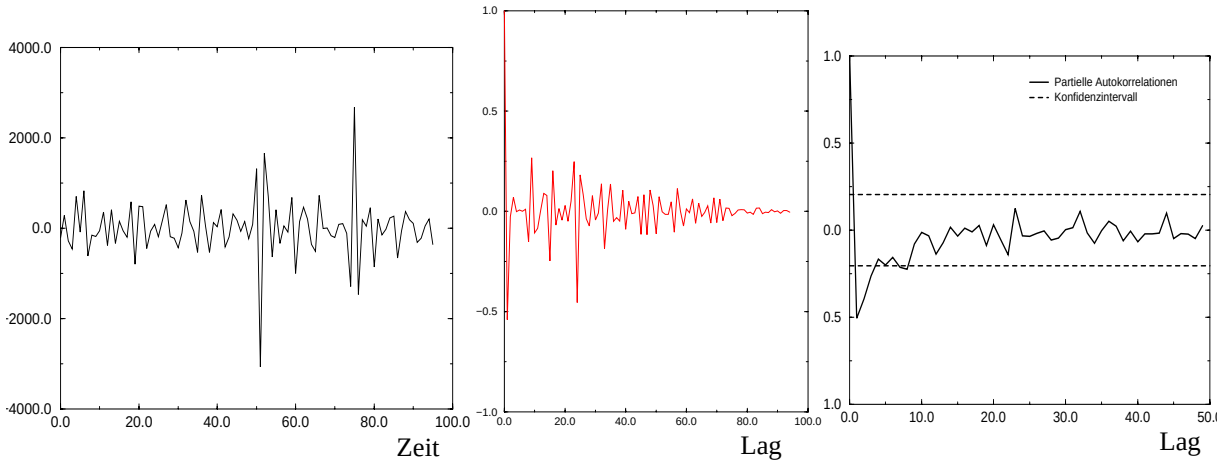


**Abbildung 3-9:** Zweite Filterung, Autokorrelationsfunktion und partielle Autokorrelationen

Die aus dieser zweiten Filterung resultierende Autokorrelationsfunktion (Abbildung 3-9 Mitte) zeigt keine der Strukturen mehr, die eine nochmalige Filterung notwendig erscheinen ließen. Abbildung 3-9 rechts zeigt außerdem noch die entsprechenden partiellen Autokorrelationen samt Konfidenzintervall $\left[\frac{-2}{\sqrt{96}}, \frac{2}{\sqrt{96}}\right]$, um im Vergleich mit Abbildung 3-3 den Effekt der Filterung auch hierauf zu demonstrieren.

### 3.3.2 Modellidentifikation

Wie bereits kurz in Abschnitt 3.1.4 erwähnt, gibt es im Falle von ARMA-Modellen kein einfaches Verfahren zur Bestimmung der Ordnung, wie dies für AR-Modelle mittels der partiellen Autokorrelationen noch möglich war. Deshalb wurden auf die Empfehlung von [SchS89] und [BJ76] hin für $p = 0, 1, 2$ und $q = 0, 1, 2$ jeweils die Modellparameter bestimmt und die Ergebnisse verglichen. Es hat sich dabei gezeigt, daß das ARMA(2,2)-Modell (mit insgesamt nur vier Parametern im Gegensatz zu den 24 Parametern beim AR-Modell) die besten Resultate vorzuweisen hatte.

### 3.3.3 Verfahren zur Parameterschätzung

Auch für die Schätzung der ARMA-Modellparameter $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ existieren mehrere Verfahren, von denen hier zwei genauer vorgestellt werden sollen.

#### (e) Durbin-Verfahren

Ein einfaches Verfahren nach [Dur60] basiert auf einem Regressionsansatz:

1. Unterstelle (vorläufig) einen AR(k)-Prozeß hoher Ordnung (etwa $k \approx \sqrt{N}$) und schätze die Parameter $\phi_1, \ldots, \phi_k$ mit Hilfe der Levinson-Durbin-Rekursion oder des Yule-Walker-Verfahrens aus Abschnitt 3.2.2.

2. Bestimmung der Residuen

$$\hat{\varepsilon}_t = x_t - \phi_1 x_{t-1} - \ldots - \phi_k x_{t-k} \qquad \text{für } t = 1, \ldots, N \tag{3.55}$$

   mit dem CLS-Ansatz aus Abschnitt 3.2.2.

3. Einsetzen der so erhaltenen Residuen in die Modellgleichung (vgl. (3.22))

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \hat{\varepsilon}_t - \theta_1 \hat{\varepsilon}_{t-1} - \theta_2 \hat{\varepsilon}_{t-2} - \ldots - \theta_q \hat{\varepsilon}_{t-q} \tag{3.56}$$

   und Berechnung der Koeffizienten aus einem Least-Squares-Ansatz, d.h. Minimierung von

$$\sum_{t=p+q+1}^{N} \left( x_t - \left( \sum_{\tau=1}^{p} \hat{\phi}_\tau x_{t-\tau} \right) - \left( \sum_{\tau=1}^{q} \hat{\theta}_\tau x_{t-\tau} \right) \right)^2 \stackrel{!}{=} \min. \tag{3.57}$$

Definiere nun analog zu (3.45)ff.

$$A := \begin{bmatrix} x_{p+q} & x_{p+q-1} & \dots & x_{q+1} & \hat{\varepsilon}_{p+q} & \dots & \hat{\varepsilon}_{p+1} \\ x_{p+q+1} & x_{p+q} & \dots & x_{q+2} & \hat{\varepsilon}_{p+q+1} & \dots & \hat{\varepsilon}_{p+2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{N-1} & x_{N-2} & \dots & x_{N-p} & \hat{\varepsilon}_{N-1} & \dots & \hat{\varepsilon}_{N-q} \end{bmatrix} \tag{3.58}$$

$$c := (\phi_1, \phi_2, \dots, \phi_p, -\hat{\theta}_1, -\hat{\theta}_2, \dots, -\hat{\theta}_q) \tag{3.59}$$

$$f := (x_{p+q+1}, x_{p+q+2}, \dots, x_N)\,. \tag{3.60}$$

Damit läßt sich (3.57) wie folgt schreiben:

$$\| f - Ac \|_2^2 \stackrel{!}{=} \min\,. \tag{3.61}$$

Das überbestimmte lineare Gleichungssystem $Ac = f$ löst man dabei wie an entsprechender Stelle in Abschnitt 3.2.2 über eine Householder-Transformation.

### (f) Marquardt-Verfahren für ARMA-Modelle

Hierbei handelt es sich um ein iteratives Minimierungsverfahren mit günstigen Konvergenzeigenschaften, das die in Abschnitt 3.2.2 (d) vorgestellte Technik für ARMA-Modelle adaptiert. Deshalb im folgenden nur die allerwesentlichsten Schritte:

1. *Bestimmung der Residuen* $\hat{\varepsilon}_t = x_t - \phi_1^{(i)} x_{t-1} - \dots - \phi_p^{(i)} x_{t-p} + \theta_1^{(i)} x_{t-1} + \dots + \theta_q^{(i)} x_{t-q}\,.$

   Als Anfangsschätzwerte $\hat{\phi}_1^{(0)}, \dots, \hat{\phi}_p^{(0)}, \hat{\theta}_1^{(0)}, \dots, \hat{\theta}_q^{(0)}$ setzt man die durch die eben beschriebene Durbin-Rekursion (e) erhaltenen Parameter ein. Wie im Fall des AR-Modells benötigt man zur Bestimmung der $\hat{\varepsilon}_t$ für $t = 1-r, \dots, 1$ auch hier nicht-beobachtbare Startwerte $x_t$ mit $t \le 0$, die man entweder mit dem dort erwähnten CLS- oder dem ULS-Ansatz erhält.

2. *Numerische Berechnung der partiellen Ableitungen* $\dfrac{\partial}{\partial \phi_k^{(i)}} \hat{\varepsilon}_t(\phi^{(i)}, \theta^{(i)})$, $\dfrac{\partial}{\partial \theta_k^{(i)}} \hat{\varepsilon}_t(\phi^{(i)}, \theta^{(i)})$ über eine $((N+r) \times (p+q))$-Matrix $D = (d_{tk})$ analog zum 2. Schritt von Verfahren (d).

3. *Lösung des überbestimmten linearen Gleichungssystems* $D^{(i)} \cdot x^{(i)} = \hat{\varepsilon}_t(\phi^{(i)}, \theta^{(i)})\,.$

   Hierzu wird statt der schlecht konditionierten Normalgleichungen wiederum die Householder-Transformation [Sto93] verwendet.

4. *Neue Schätzwerte für die Modellparameter nach der* $(i+1)$*-ten Iteration:*

$$(\hat{\phi}_1^{(i+1)}, \dots, \hat{\phi}_p^{(i+1)}, \hat{\theta}_1^{(i+1)}, \dots, \hat{\theta}_q^{(i+1)}) = (\hat{\phi}_1^{(i)}, \dots, \hat{\phi}_p^{(i)}, \hat{\theta}_1^{(i)}, \dots, \hat{\theta}_q^{(i)}) - (x_1^{(i)}, \dots, x_{p+q}^{(i)}) \tag{3.62}$$

5. *Rücksprung zu Iterationsschritt 1.*

Das Abbruchkriterium hat auch hier die Form $max\left|x_k^{(i)}\right| < \varepsilon$ für kleines $\varepsilon$. Als Schätzer für die Residualvarianz erhält man schließlich

$$\hat{\sigma}_\varepsilon^2 = \frac{S^2(\hat{\phi}_1^{(i)}, \ldots, \hat{\phi}_p^{(i)}, \hat{\theta}_1^{(i)}, \ldots, \hat{\theta}_q^{(i)})}{N-(p+q)}. \tag{3.63}$$

### 3.3.4 Resultat und Diagnose

Auch im Fall der ARMA-Modellierung wurden beide vorgestellte Verfahren implementiert; im folgenden wird exemplarisch auf das durch die Marquardt-Technik erhaltene Ergebnis eingegangen. Die Parameterschätzung erbrachte das in Tabelle 3-10 angegebene Resultat.

| | **Marquardt (f)** |
|---|---|
| $\hat{\phi}_1$ | 0.431459 |
| $\hat{\phi}_2$ | -0.016839 |
| $\hat{\theta}_1$ | 1.609863 |
| $\hat{\theta}_2$ | -0.651051 |

**Tabelle 3-10:** Parameterschätzung für das ARMA(2,2)-Modell der gefilterten Referenzreihe

Mit dem Diagnose-Verfahren von Satz 3.14 läßt sich zunächst die Güte der Anpassung des ARMA(2,2)-Modells an die durch zweimalige Filterung erhaltene Zeitreihe $z_t$ (vgl. (3.54)) überprüfen. Für die Box-Pierce-Statistik mit $m = \sqrt{96} \approx 10$ und $p = q = 2$ erhält man $Q = 10.88$. Der Vergleich mit einer $\chi^2[6]$-Verteilung liefert dafür ein $(\alpha = 0.9)$-Quantil. Dies ist ein starker Hinweis darauf, daß die doppelt gefilterte Reihe von Abbildung 3-9 tatsächlich als Realisierung eines ARMA(2,2)-Prozesses angesehen werden kann.

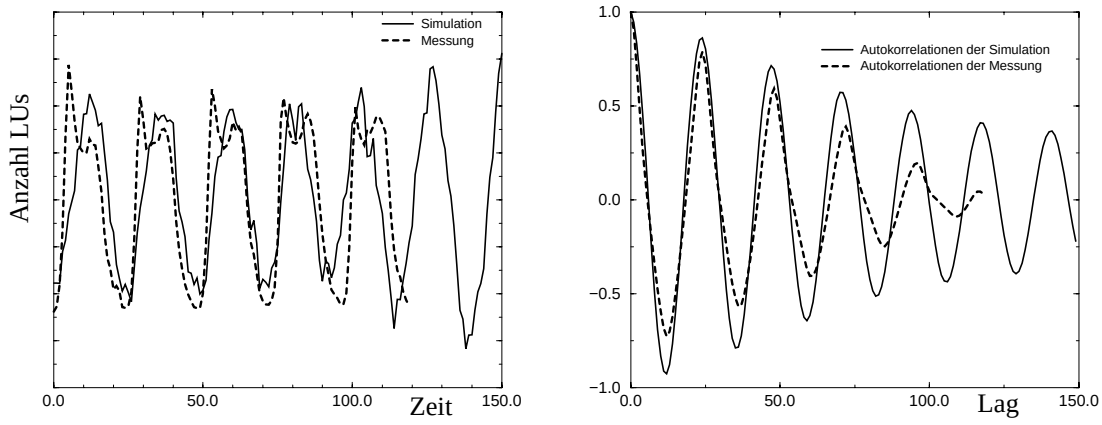


**Abbildung 3-11:** Das ARMA(2,2)-Modell der Referenzreihe und zugehörige Autokorrelationsfunktionen

Abbildung 3-11 zeigt abschließend das ARMA(2,2)-Modell der Referenzreihe, das aus dem Modell für die Zeitreihe $z_t$ durch Rückgängigmachung der beiden Filterungstransformationen (3.54) und (3.53) erhalten wurde. Schon der Augenschein zeigt allerdings, daß trotz des guten Ergebnisses der Diagnose und des guten Eindrucks beim Vergleich von Referenz- und Modellreihe selbst (Abbildung 3-11 links) das ARMA-Modell doch noch immer signifikant von der vorgegebenen Autokorrelationsfunktion abweicht; zudem schwingt sich die Autokorrelationsfunktion des ARMA-Modells mit zunehmender Zeit auf eine Periode von etwa 23 anstatt der erwünschten 24 ein.

## 3.4 Fazit

Dieses Kapitel war dem Versuch gewidmet, das in Kapitel 2 formulierte Referenzproblem mit klassischen Modellierungsansätzen anzugehen. Nachdem sich die Klasse der autoregressiven Prozesse als möglicherweise geeigneter Kandidat herausgestellt hatte, erfolgte eine detaillierte Einführung in die mathematischen Grundlagen der AR(p)- und ARMA(p,q)-Prozesse sowie den Ansatz von Box-Jenkins zur Modellidentifikation und -parametrisierung einschließlich der zugehörigen Standardverfahren. Die anschließend durchgeführte Modellierung der Referenzkurve verlief jedoch leider in allen Fällen unbefriedigend. Zudem stellte sich heraus, daß aufgrund des Aufwands für die Parametrisierung autoregressive Prozesse für den Einsatz in GSM-Systemtests kaum in Frage kommen. Vor diesem Hintergrund stellt sich erneut die Aufgabe, ein einfaches Modellierungsverfahren zu entwickeln, das dennoch die wesentlichen statistischen Charakteristiken der Referenzkurve wie Randverteilung, Autokorrelationsfunktion und Verlauf hinreichend nachmodelliert. Die im folgenden Kapitel vorgestellte TES-Methode und ihre Erweiterung werden sich als eine mögliche Lösung dieses Problems herausstellen.

KAPITEL 4

# *Die TES-Methode und ihre Verallgemeinerung*

## 4.1 Grundlagen und Standardverfahren

Nachdem die Ergebnisse von Kapitel 3 gezeigt haben, daß eine Behandlung des Referenzproblems mit autoregressiven Modellen den besonderen Anforderungen eines Einsatzes innerhalb einer GSM-Testumgebung nur bedingt gerecht werden kann, widmet sich das vorliegende Kapitel der Weiterentwicklung des sogenannten TES-Modells[1], das sich - wie wir sehen werden - besser dafür eignet. Nach der Einführung in das Standardverfahren und seine Parametrisierung entwickeln wir in Abschnitt 4.3 das neue Konzept der “Generalized Stitching Function” und demonstrieren es an einigen Beispielen, bevor in Abschnitt 4.4 noch ein Ansatz zur automatisierten Parametrisierung von TES-Modellen vorgeschlagen und validiert wird.

### 4.1.1 Anforderungen an das TES-Verfahren: die “Drei Kriterien”

Das Standard-TES-Verfahren, wie es von Jagerman und Melamed ausgearbeitet wurde (vgl. [JM92a], [JM92b], [Mel93]), wurde als Verfahren zur Modellierung von Zeitreihen im Hinblick auf die Erfüllung folgender drei Kriterien entwickelt:

(i) Die Randverteilung der modellierten Kurve soll exakt einer vorgegebenen Randverteilung entsprechen. Falls insbesondere eine vorgegebene empirische Zeitreihe im Modell nachgebildet werden soll, haben die Randverteilungen beider Kurven übereinzustimmen.

1. Die Abkürzung TES steht für “Transform - Expand - Sample” und leitet sich von der Art und Weise her, auf die im ersten Schritt der einfachsten Version des Verfahrens die sogenannte “Innovationssequenz” (vgl. Abschnitt 4.1.2) gewonnen wird. Dies läßt sich folgendermaßen veranschaulichen: Auf einem Kreis von Umfang 1 springt man (ausgehend vom letzten Glied der Sequenz) um einen fixen Betrag $\varphi$ weiter (“Transform”), errichtet um den erreichten Wert ein symmetrisches Intervall mit vorher festgelegter Länge $\delta$ (“Expand”) und würfelt sodann gleichverteilt einen Wert aus diesem Intervall (“Sample”), der als nächster Wert in die Innovationssequenz aufgenommen wird. In Abschnitt 4.1.2 wird dieses Vorgehen detaillierter eingeführt, insbesondere werden wir sehen, daß sich $\varphi$ und $\delta$ über die sogenannte “Innovationsdichte” ausdrükken lassen, hierbei entspricht $\varphi$ dem Mittelwert dieser Dichte und $\delta$ der Breite ihres Trägerintervalls.

(ii) Die Autokorrelationsfunktion der modellierten Zeitreihe soll zumindest in den führenden Autokorrelationen eine gegebene Funktion (insbesondere ggf. die Autokorrelationsfunktion einer gegebenen empirischen Zeitreihe) hinreichend genau approximieren.

(iii) Schließlich soll die modellierte Zeitreihe der vorgegebenen auch visuell so weit wie möglich ähneln. Hierbei handelt es sich offenbar um ein eher qualitativ-subjektives Kriterium.

Auf einer sehr abstrakten Ebene besteht das TES-Verfahren in der Reduktion eines linearen autoregressiven Prozesses auf das Einheitsintervall, worauf noch einige zusätzliche Transformationen ausgeführt werden. Letztere resultieren im Verlust der Markoveigenschaft. Klassifizieren läßt sich TES nach [LM82] allgemein als approximierende Korrelations-Verzerrungs-Methode.

Vorab lassen sich folgende entscheidende Vorteile eines TES-Modells festhalten:

- Ein TES-Modell erfüllt Kriterium (i) exakt, Kriterium (ii) hinreichend genau, und bezüglich der visuellen Ähnlichkeit von Kriterium (iii) lassen sich bei geeigneter Parametrisierung durchaus sehenswerte Ergebnisse erzielen.
- Eine TES-Sequenz ist sehr leicht zu generieren, Zeit- und Speicherkomplexität sind praktisch vernachlässigbar
- Es gibt eine (wenngleich nicht triviale) Möglichkeit, die Autokorrelationsfunktion einer TES-Sequenz analytisch anhand ihrer Parametrisierung zu berechnen. Dies macht zumindest in der Theorie eine simulationsbasierte Schätzung der Autokorrelationen überflüssig.

Im folgenden wird nun kurz der grundsätzliche Aufbau der Standard-TES-Methode zusammengefaßt. Jagerman und Melamed unterscheiden grundsätzlich zwei verschiedene Arten von TES, die sich anhand der Autokorrelation zum Lag 1 differenzieren lassen: Ist diese Autokorrelation negativ, so ist in aller Regel das $TES^-$ genannte Verfahren besser geeignet, während für positive Lag-1-Autokorrelationen $TES^+$ zu verwenden ist. Da wir in dieser Arbeit das TES-Verfahren vor allem im Hinblick auf die Modellierung periodischen Verkehrs untersuchen und dieser in allen typischen Fällen eine positive Lag-1-Autokorrelation aufweisen wird (z. B. weist unsere Referenzmessung eine Periode von 24 auf, d.h. die Autokorrelationsfunktion bleibt etwa bis Lag 6 positiv, vgl. Abbildung 2-4 links), werden wir auf die $TES^-$-Variante nicht weiter eingehen und daher die $TES^+$-Variante kurzerhand als TES-Methode bezeichnen.

### 4.1.2 Die Ebenen des TES-Schemas

Abbildung 4-1 stellt die Grundstruktur der TES-Methode schematisch dar. Im wesentlichen lassen sich hierbei folgende Ebenen unterscheiden:

- **Initialisierung:** Hier findet die Initialisierung des gesamten Schemas durch Würfeln der Zufallsvariablen $Z_i$ für jeden Schritt $i = 0, 1, 2, \ldots$ aus einer (0,1)-Gleichverteilung statt.
- **Innovationssequenz:** In dieser Ebene werden die zufällig gewählten $Z_i$ für $i \geq 1$ dazu verwendet, aus einer beliebigen über dem Einheitsintervall definierten Verteilung, der sogenannten "*Innovationsverteilung*", die Zufallsvariable $V_i$ zu bestimmen. Die sog. Innovationssequenz $(V_i)_{i = 1, 2, \ldots}$ konzentriert sich dabei für gegebenes $Z_0$ auf den Träger der entsprechenden Innovationsdichte (Abbildung 4-2 links).
- **Hintergrundsequenz:** Auf dieser Ebene werden Korrelationen zwischen den einzelnen Sequenzmitgliedern erzeugt. Dies geschieht dadurch, daß $Y_{i-1}$, also gewissermaßen das "bisher letzte" Element der Hintergrundsequenz, mit $V_i$ zusammengezählt wird, und zwar mittels Modulo-1-Arithmetik:

  $$Y_i = \langle Y_{i-1} + V_i \rangle_{\text{mod } 1}. \tag{4.1}$$

  Wir werden später sehen, daß die so entstandene Hintergrundsequenz im Fall einer nicht zum Ursprung (modulo 1) symmetrischen Innovationsdichte im wesentlichen mit fester Periode um einen Kreis mit Umfang 1 herumwandert.

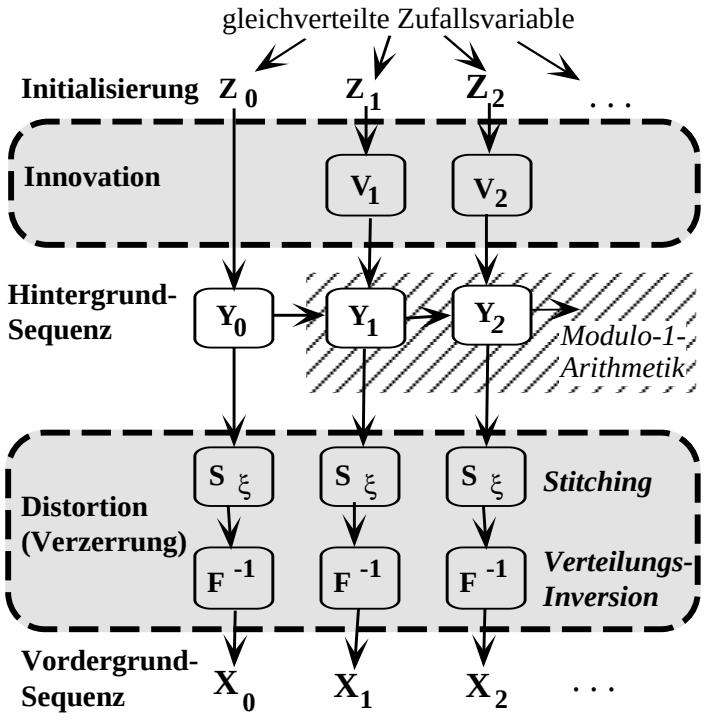


**Abbildung 4-1:** Das Schema des Standard-TES-Verfahrens

- **Verzerrung (Distortion):** Die Hintergrundsequenz erfährt - wie bereits erwähnt - noch zwei Transformationen. Zum einen deshalb, weil die Modulo-1-Rechnung aus dem vorigen Abschnitt zu unerwünschtem Sprungverhalten führen kann: Wenn etwa zu einer Hintergrund-Zufallsvariablen $Y_i$, die ganz knapp unterhalb von 1 liegt, eine nur sehr kleine Innovationsvariable $V_{i+1}$ addiert wird, kann die Operation im Endergebnis zu einem Sprung von 1 nach 0, also einer unverhältnismäßig großen Änderung bzw. diskontinuierlichem Verhalten der Hintergrundsequenz führen. Aus diesem Grunde wurde die sogenannte “Stitching-Transformation” eingeführt[2], die derartige Sprünge glättet. Die Standard-Stitching-Funktion (SSF) hat dabei (parametrisiert durch den sog. Stitching-Parameter $\xi \in [0, 1]$) den in Abbildung 4-2 rechts gezeigten stückweise linearen Verlauf

$$\xi(y) = \begin{cases} \dfrac{y}{\xi} & \text{falls } 0 \le y \le \xi \\ \dfrac{1-y}{1-\xi} & \text{falls } \xi \le y < 1 \end{cases} \tag{4.2}$$

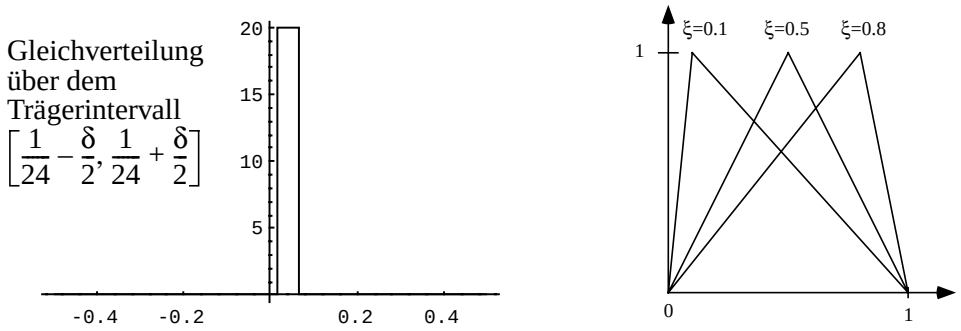


**Abbildung 4-2:** Einfachster Fall einer Innovationsdichte (links) und Standard-Stitching-Funktionen

Abbildung 4-3 zeigt exemplarisch, daß die Sprünge der Hintergrundsequenz (Mitte) nach Anwendung der SSF geglättet sind (rechts).

Dabei ist festzuhalten, daß sich an der Randverteilung der beiden Sequenzen durch die Stitching-Transformation nichts ändert, was daran liegt, daß die SSF – selbst aufgefaßt als eine Zeitreihe – als Randverteilung die (0,1)-Gleichverteilung aufweist (vgl. Definition A.4). Dies ist deshalb so wichtig, weil sich beweisen läßt (vgl. Anhang A Lemma A.1), daß auch die Hintergrundsequenz $(Y_i)_{i=1,2,\ldots}$ auf dem Einheitsintervall gleichverteilt ist und damit auch die gestitchte Hintergrundsequenz. Dies führt zu

**Bemerkung 4.1:** Beim Standard-TES-Verfahren besitzt auch die gestitchte Hintergrundsequenz als Randverteilung die (0,1)-Gleichverteilung.

2. Der Name leitet sich von engl. to stitch = zusammenheften ab.

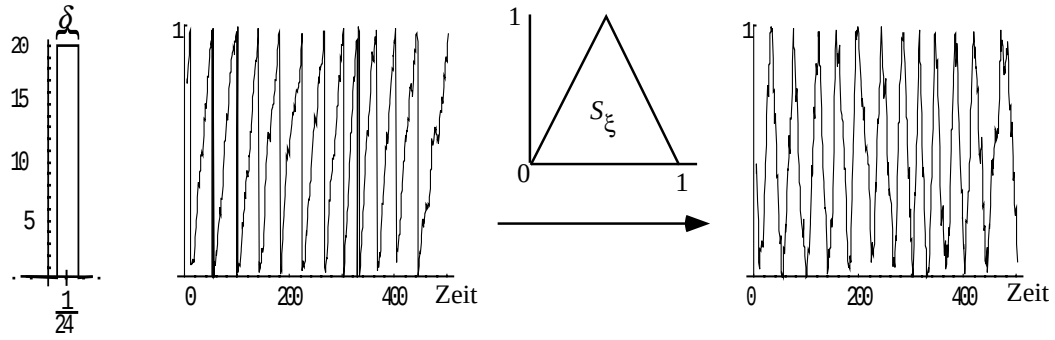


**Abbildung 4-3:** Exemplarische Innovationsdichte, Hintergrundsequenz, SSF und gestitchte Vordergrundsequenz

Diese Tatsache bildet die Voraussetzung für die Wirksamkeit der zweiten Verzerrungstransformation, nämlich der Histogramm-Inversion. In Abbildung 2-4 rechts sind wir schon der Randverteilung begegnet, die die modellierte Zeitreihe letztendlich aufweisen soll. Lemma A.3 aus Anhang A stellt uns eine Transformation zur Verfügung, die durch Anwendung der invertierten Randverteilung auf die gestitchte Hintergrundsequenz eine neue Sequenz erzeugt, die exakt die vorgegebene Randverteilung aufweist. Diese sog.

- **Vordergrundsequenz** stellt schlußendlich das Resultat der Modellierung dar und kann in unserer Umgebung dann etwa als Verkehrsreihe für den GSM-Systemtest verwendet werden.

Soweit ein kurzer Überblick über die Elemente des Standard-TES-Schemas. Für eine detaillierter Beschreibung sei nochmals auf die genannte Literatur (vor allem [JM92a], [JM92b], [Mel93]) verwiesen. Im Anhang A finden sich zudem einige Lemmata, die für diese Untersuchungen von entscheidender Wichtigkeit sind, samt ihren Beweisen dargestellt. Im folgenden Abschnitt wird noch kurz auf die Berechnung der Autokorrelationsfunktion einer TES-Vordergrundsequenz eingegangen, um insbesondere auf einige kleinere Ungenauigkeiten in [Mel93] hinzuweisen.

### 4.1.3 Die Autokorrelationsfunktion

Eine der ansprechenden Eigenschaften der TES-Methode beruht auf der Möglichkeit, die Autokorrelationsfunktion eines Modells in geschlossener Form anzugeben. Im Anhang A.2 wird dies an einem einfachen Beispiel detailliert veranschaulicht, während an dieser Stelle nur kurz auf das allgemeine Ergebnis eingegangen werden soll.

Ausgehend von einer Betrachtung der Hintergrundsequenz $Y_n$ als stationären Markovprozeß lassen sich dessen Übergangsdichten berechnen und mit Hilfe der Laplace-Transformierten der Innovationsdichte $f_V$ ausdrücken. Dies führt nach längerer Rechnung auf folgenden Ausdruck für die Autokorrelationsfunktion $R_X(\tau)$ einer TES-Vordergrundsequenz $X$ (abhängig vom

Lag $\tau$) mit allgemeiner Distortion-Funktion $D$ (d.h. Stitching-Transformation *und* Histogramm-Inversion) bzw. deren Laplacetransformierter $\tilde{D}$:

$$R_X(\tau) = \frac{2}{\sigma_X^2} \sum_{\nu=1}^{\infty} Re[(\tilde{f}_V(2\pi i\nu))^{\tau}] \left|\tilde{D}(2\pi i\nu)\right|^2 \tag{4.3}$$

An dieser Stelle ist die Darstellung in [Mel93] insofern ungenau, als hier - ohne dies ausdrücklich zu erwähnen - die sonst übliche Definition der Laplacetransformierten

$$\tilde{D}(s) = \int_0^{\infty} e^{-st} D(t)dt \tag{4.4}$$

ersetzt wird durch

$$\tilde{D}_{JM}(s) = \int_0^{1} e^{-st} D(t)dt \tag{4.5}$$

also durch die Laplacetransformation der auf das Einheitsintervall restringierten Distortion-Funktion.

Die numerische Problematik von Formel (4.3) hängt offensichtlich mit der Frage der Konvergenz der unendlichen Summe zusammen. Deshalb wurde diese für den Fall einer Innovationsdichte der in Abbildung 4-2 links vorgestellten Form exemplarisch untersucht; das Ergebnis ist für Lags 1, 2, 5, 10 und 20 in Abbildung 4-4 dargestellt.

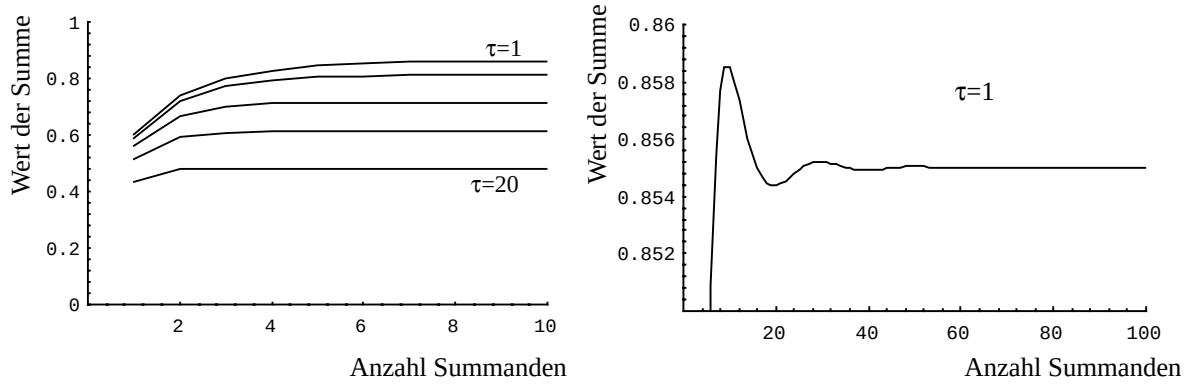


**Abbildung 4-4:** Zur Konvergenz von (4.3) in Abhängigkeit von der Anzahl der einbezogenen Summanden: links die Darstellung für Lags $\tau = 1, 2, 5, 10, 20$ (der Reihe nach von oben), rechts die Entwicklung für $\tau = 1$ in höherer Auflösung

Es bleibt festzuhalten, daß die fragliche unendliche Summe in der Tat sehr schnell konvergiert. Die linke Abbildung legt den Schluß nahe, daß für alle Lags ein quasi monotoner Anstieg der

Summe (in Abhängigkeit von der Anzahl der in die Summe einbezogenen Summanden) vorliegt, der schon für eine relativ kleine Anzahl (Größenordnung 7) hinreichend konvergiert. Bei einer um den Faktor 1000 höheren Auflösung stabilisiert sich (Abbildung 4-4 rechts) die Summe im "worst case" $\tau = 1$ ebenfalls nach höchstens 20 Gliedern[3] so, daß es für alle praktischen Anwendungen ausreicht. Vor diesem Hintergrund erweist sich der Ausdruck (4.3) als numerisch stabil und unproblematisch. Die Frage der Berechnung von Laplacetransformierten für nicht-triviale Distortion-Funktionen bleibt davon natürlich unberührt.

Abschließend eine Bemerkung zur Frage, wie gut denn simulative (d.h. aus dem TES-Modell empirisch ermittelte) und numerische (d.h. aus der analytischen Formel hervorgehende) Autokorrelation übereinstimmen. Die Antwort von [Mel93] ist in Abbildung 4-5 dargestellt, in der behauptet wird, daß bereits eine Simulation der Länge 500 Resultate erbringe, die identisch mit dem entsprechenden analytischen Ergebnis sind.

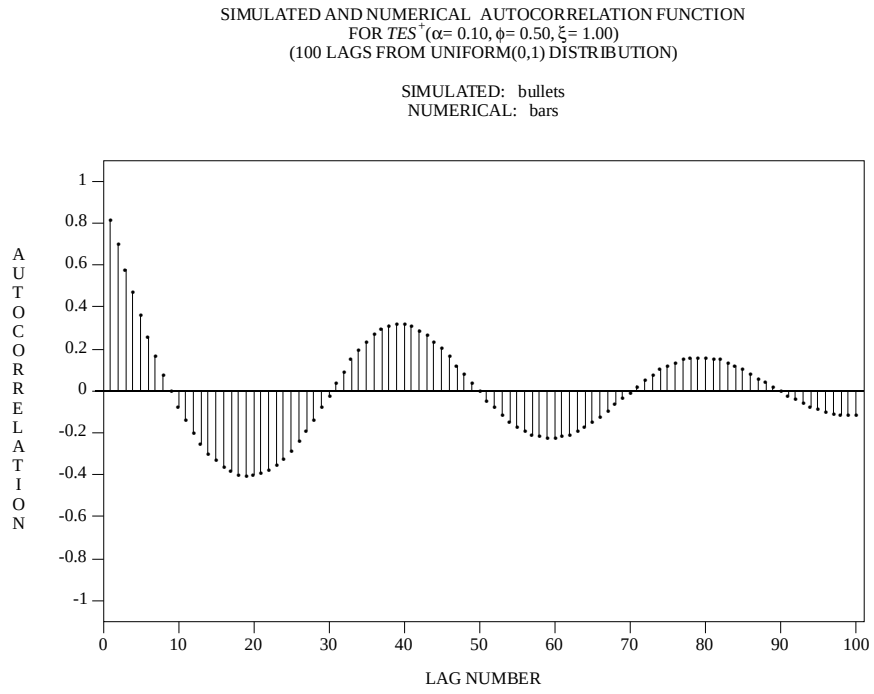


**Abbildung 4-5:** Zur Übereinstimmung von analytischer und empirischer Autokorrelation nach Melamed (entnommen aus [Mel93])

Eigene Untersuchungen konnten diese Aussage nur bedingt bestätigen: Abbildung 4-6 zeigt das Simulationsergebnis für ein identisch zu [Mel93] parametrisiertes Modell. Hieraus läßt sich ersehen, daß für eine Simulationslänge von 500 bei weitem nicht die behauptete Übereinstimmung von analytischer (fett) und simulativer (normal) Autokorrelationsfunktion erzielt

3. Nach 20 Iterationen ergibt sich der Summenwert zu 0.855 ± 0.000615, nach 30 Iterationen zu 0.855 ± 0.000202, nach 40 Iterationen zu 0.855 ± 0.000089 und nach 50 Iterationen zu 0.855 ± 0.000047. Konvergenz auf 0.855 ± 0.000001 wird nach ca. 150 Iterationen erreicht.

wurde; erst bei einer Simulationslänge von 10 000 (rechte Kurve) kann man von guter Übereinstimmung sprechen.

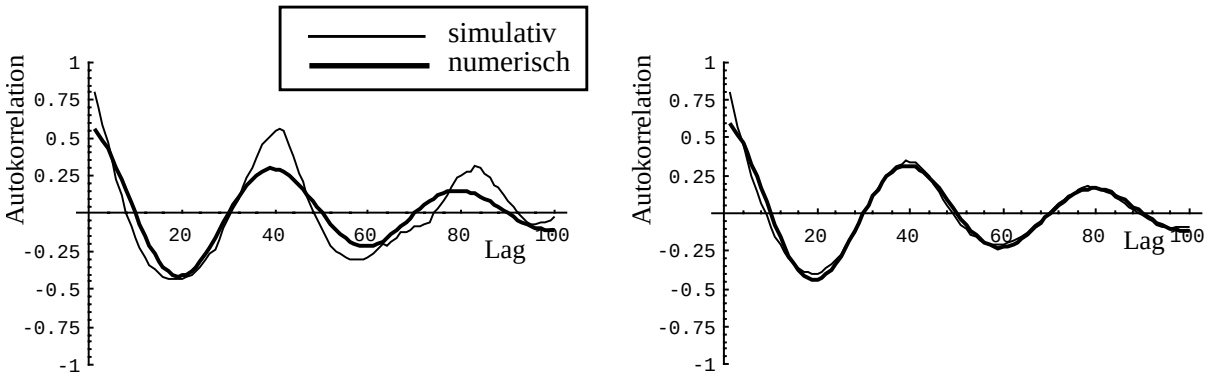


**Abbildung 4-6:** Übereinstimmung von analytischer und simulativer Autokorrelation: eigenes Ergebnis für das gleiche Modell wie in Abbildung 4-5 (Simulationslänge links 500, rechts 10 000)

## 4.2 Parametrisierung und Resultate

Nachdem der letzte Abschnitt in die wesentlichen Bausteine der Standard-TES-Methode eingeführt hat, soll nun kurz auf die Bedeutung der einzelnen Parameter eingegangen werden. Ein TES-Modell ist charakterisiert durch

- Initialisierungswert,
- Innovationsdichte und
- Stitching-Funktion, d.h. im Standardfall Stitching-Parameter.

Hiervon hat die Wahl des Initialisierungswertes keine größere Bedeutung für die Güte des Modells [JM92a]. Mit der Innovationsdichte verhält es sich da schon anders. Sie beeinflußt insbesondere Periodizität und Varianz des resultierenden Modells.

Grundsätzlich kann jede Verteilungsfunktion als Innovationsverteilung herangezogen werden. In der Praxis hat sich jedoch der Vorschlag von [Mel93] bewährt, sich auf Innovationsdichten der Form

$$f(x) = \sum_{i=1}^{N} \frac{p_i}{b_i - a_i} 1_{[a_i, b_i)} \tag{4.6}$$

zu beschränken, die stückweise gleichverteilt sind. Der Vorteil bei dieser Klasse von Verteilungsfunktionen liegt in ihrer großen Einfachheit und zugleich Flexibilität, da sich mit ihrer Hilfe grundsätzlich jede Verteilung beliebig genau annähern läßt. Zudem haben die Untersu-

chungen in [E-H97] gezeigt, daß sich (abgesehen von einer möglichen Parameter-Reduktion) kein systematischer Vorteil bei Verwendung anderer weitverbreiteter Verteilungen ableiten ließ.

Die Periodizität einer Vordergrundsequenz hängt direkt vom Schwerpunkt der Innovationsverteilung ab. Liegt dieser im Ursprung, so nennt man das resultierende Modell “driftless”, und die Hintergrundsequenz hat beispielsweise die Gestalt von Abbildung 4-7 links.

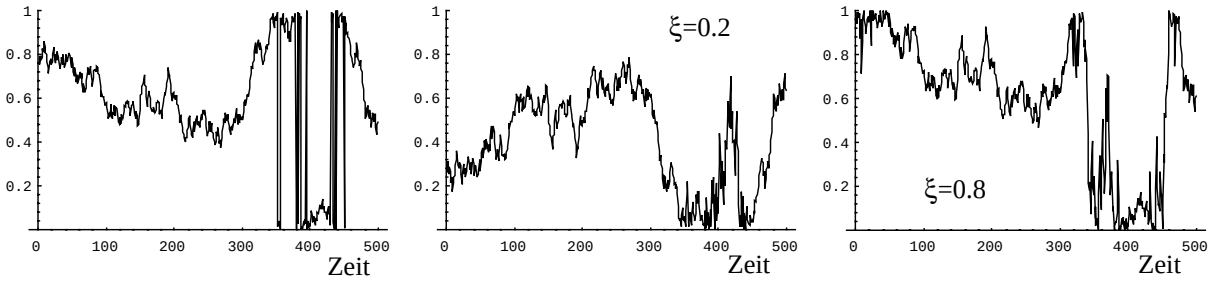


**Abbildung 4-7:** Driftlose Hintergrundsequenz (links Unstetigkeiten vor dem Stitching, Mitte Standard-Stitching-Transformation mit Parameter 0.2 bzw. 0.8).[4]

An diesem Beispiel läßt sich zugleich der Einfluß des Stitchingparameters studieren. Wir sehen (Abbildung 4-7 Mitte und rechts), daß unterschiedliche Stitchingparameter (hier $\xi = 0.2$ bzw. $\xi = 0.8$ zwar beide Male die Unstetigkeiten der ungestitchten Kurve glätten, aber doch zu sehr unterschiedlichen Sequenzen führen können.

Liegt der Schwerpunkt der Innovationsdichte dagegen nicht im Ursprung, so zeigt die resultierende Hintergrundsequenz eine Periodizität. Abbildung 4-8 zeigt zwei TES-Modelle mit einfachen Innovationsdichten gemäß (4.6) (mit $N = 1$), deren Schwerpunkt jeweils bei $\frac{1}{24}$ liegt. Die resultierende Hintergrund- (und damit auch letztlich die Vordergrundsequenz) erhält dadurch eine Periode von 24. Dieser Typ von TES-Modellen wird uns im weiteren ausschließlich beschäftigen.

Abbildung 4-8 zeigt auch exemplarisch den Einfluß der Breite des Trägers der Innovationsdichte insbesondere auf die Autokorrelationsfunktion. Diese Breite beträgt in der oberen Reihe $\delta = 0.01$ (d.h. das entsprechende abgebildete Intervall ist $[1/24 - 0.005, 1/24 + 0.005]$) und hat einen “ziemlich deterministischen” Verlauf der Hintergrundsequenz (Mitte) zur Folge, was sich auch in der rechts abgebildeten Autokorrelationsfunktion ausdrückt. Dagegen bewirkt eine Verfünffachung der Breite (untere Reihe), also $\delta = 0.05$, bei gleichem Schwerpunkt zwar nur eine geringfügige Änderung im Verlauf der Hintergrundsequenz, aber hat doch erhebliche Auswirkungen auf die Autokorrelationsfunktion (unten rechts).

4. NB: Die Werte der Hintergrundsequenz (Y-Achse) liegen per definitionem immer in [0,1).

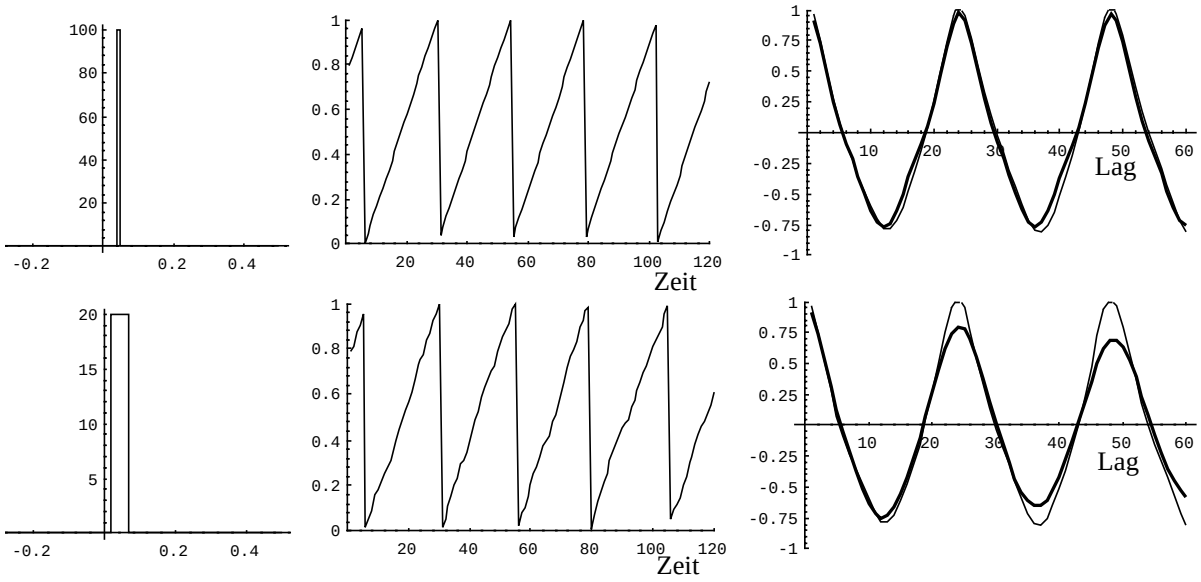


**Abbildung 4-8:** Einfache Innovationsverteilungen, driftende Hintergrundsequenzen und zugehörige Autokorrelationsfunktionen (TES-Modell fett, Referenzreihe normal abgebildet)

Mit Hilfe der letzteren Innovationsdichte ergibt sich Abbildung 4-9 als Vordergrundsequenz des Standard-TES-Modells für die Referenzkurve (gestrichelt) von Abbildung 2-3. Hierbei wurde eine SSF mit Stitchingparameter $\xi = 0.8$ verwendet.

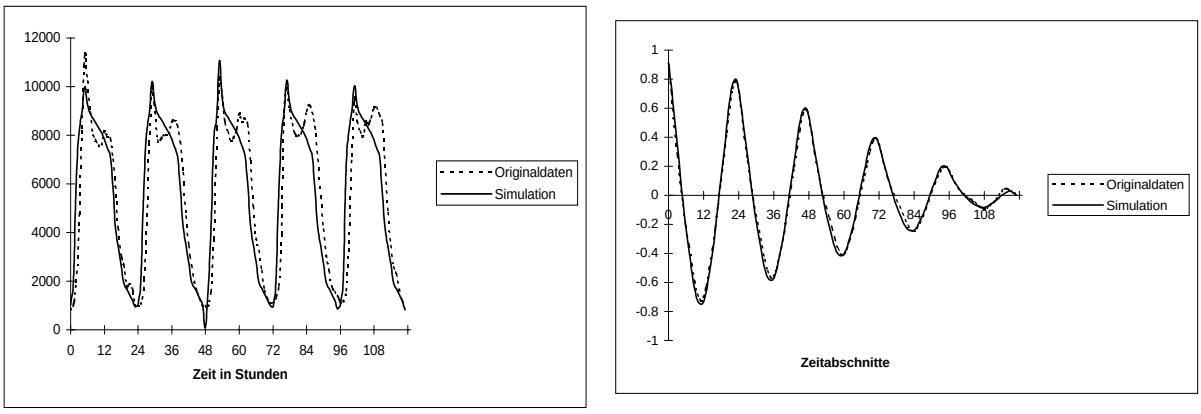


**Abbildung 4-9:** Standard-TES-Modellierung der Referenzkurve (Abbildung 2-3) und Autokorrelationsfunktion

Dieses Modell erfüllt zwar die eingangs konstatierten Kriterien (i) und (ii) hinreichend gut (insbesondere stimmt nach Abbildung 4-9 links die Autokorrelationsfunktion der modellierten Zeitreihe gut mit der vorgegebenen Autokorrelationsfunktion überein), läßt aber hinsichtlich der visuellen Ähnlichkeit zur Referenzkurve (in Abbildung 4-9 links gestrichelt dargestellt) doch zu wünschen übrig; insbesondere ist das charakteristische Mittagsloch nicht nachzumo-

dellieren. Daran ändert auch die Verwendung einer komplexeren Innovationsdichte nicht viel: Abbildung 4-10 zeigt die Vordergrundsequenzen dreier TES-Modell mit Standard-Stitching-Funktion (diesmal mit $\xi = 0.5$) und Innovationsdichten, die (bei gleichem Schwerpunkt) aus 1, 2 oder 3 Trägerintervallen (gem. (4.6)) aufgebaut sind.

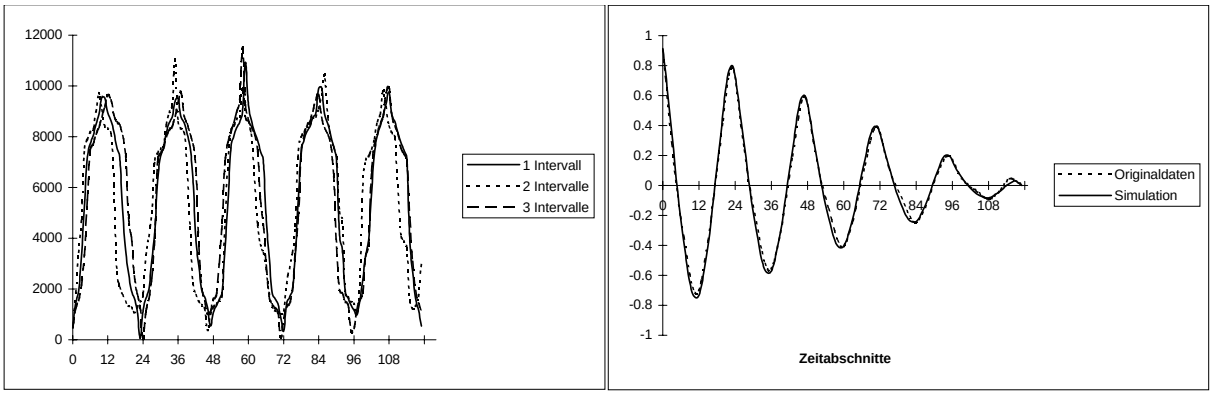


**Abbildung 4-10:** Typisches Aussehen von Standard-TES-Modellen für die Referenzkurve und zugehörige Autokorrelationsfunktionen. Es wurden geeignete Innovationsdichten der Form (4.3) mit Anzahl der Intervalle $N = 1, 2, 3$ verwendet.

Aufgrund dieser Ergebnisse wurde es notwendig, das TES-Konzept weiterzuentwickeln, und zwar erwies sich die Stitching-Funktion als richtiger Ansatzpunkt für eine Verbesserung des Modells, wie im folgenden Abschnitt 4.3 dargestellt wird.

## 4.3 Generalized Stitching Function GSF: Idee, Konsequenzen, Ergebnisse

### 4.3.1 Der Einfluß der Stitching-Funktion und die Idee einer Verallgemeinerung

Will man das TES-Verfahren, wie wir es bisher kennengelernt haben, in einem Satz zusammenfassen, so läßt sich sagen, daß die Innovationsvariablen für das Einbringen einer Zufallskomponente verantwortlich sind, die Hintergrundsequenz bringt Autokorrelationen ins Spiel, die Stitching-Funktion glättet auftretende Sprünge, und die Histogramminversion sorgt schließlich dafür, daß die entstandene Vordergrundsequenz der vorgegebenen Randverteilung genügt, während sich die Autokorrelationsfunktion aus der Kombination aller Schritte ergibt.

In diesem Abschnitt soll nun der Einfluß der Stitching-Funktion genauer untersucht werden. Sehr schnell stellt sich nämlich heraus [Rei98a], daß dieser weit über eine bloße Glättung hinausreicht.

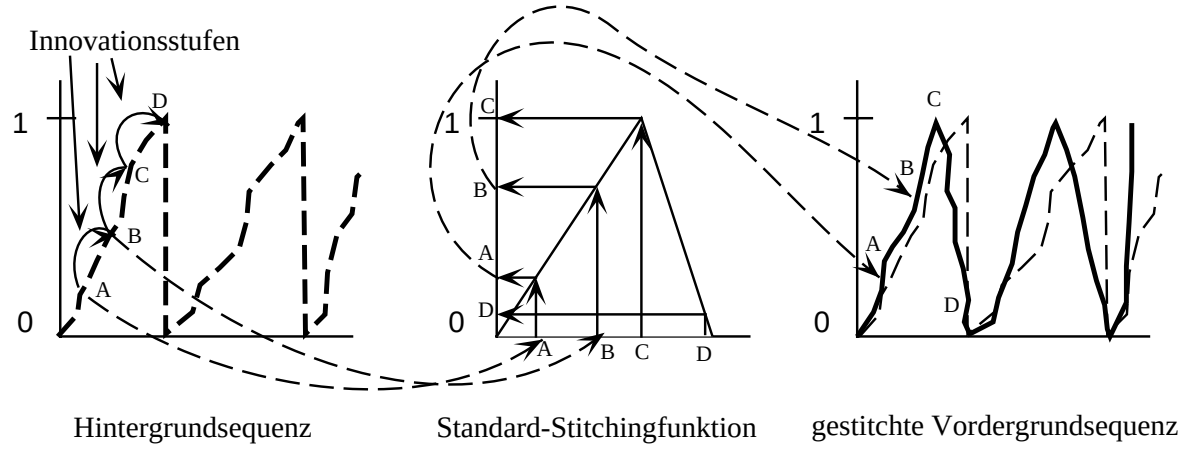


**Abbildung 4-11:** Einfluß der Stitching-Funktion

Abbildung 4-11 zeigt noch einmal genauer, was bei der Anwendung der Standard-Stitching-Funktion (SSF) gem. Gleichung (4.2) auf eine driftende Hintergrundsequenz geschieht. Links sehen wir, wie sich die Hintergrundsequenz aufgrund der Innovationsvariablen stückweise aufbaut und wegen der Modulo-1-Arithmetik bei Erreichen der 1 nach 0 zurückspringt. Die Anwendung der SSF (Abbildung 4-11 Mitte und rechts) bewirkt, daß Werte der Hintergrundsequenz, die in der oberen Region, also nahe der 1, liegen (wie dies etwa für Punkt D der Fall ist), auf kleine Werte (nahe bei 0) abgebildet werden. Nach dem Stitching liegt somit eine geglättete Sequenz vor, deren Gestalt aber in jeder einzelnen Periode als eine verzerrte Version der SSF selbst gedeutet werden kann (in unserem Beispiel also im Verlauf des Einheitsintervalls mehr oder weniger unruhig zu einem Maximum bei 1 emporsteigt und danach wieder zu 0 zurückläuft). Abbildung 4-9 und 4-10 demonstrieren weiterhin, daß auch nach Anwendung der Histogramminversion die TES-Vordergrundsequenzen immer noch im wesentlichen die Gestalt der SSF aufweisen.

Diese Beobachtung führt geradewegs auf die Idee einer Verallgemeinerung der Stitching-Funktion ([Rei98a], [Rei98b]). Wenn die Standard-Stitching-Funktion die Gestalt der Vordergrundsequenz derart maßgeblich mitbestimmt, so sollte es umgekehrt möglich sein, aus einer gewünschten Gestalt der Vordergrundsequenz eine allgemeinere Funktion abzuleiten, die dann als “Generalized Stitching Function” GSF die Stelle der SSF einnimmt, als solche auf die Hintergrundsequenz angewendet wird und letztendlich eine Vordergrundsequenz der gewünschten Gestalt hervorbringt. Wie sich im folgenden Abschnitt zeigt, ist die Bestimmung einer solchen GSF nicht schwierig.

### 4.3.2 Die Bestimmung einer Verallgemeinerten Stitching-Funktion

Zunächst beschreiben wir die Anforderungen an eine Verallgemeinerung der Stitching-Funktion formal:

**Definition 4.2:** *Eine stetige Funktion* $S: [0, 1) \to [0, 1)$ *heißt Verallgemeinerte Stitching-Funktion (Generalized Stitching Function GSF), wenn folgende Bedingungen erfüllt sind:*

- $S(0) = 0$
- $\lim_{x \to 1} S(x) = 0$
- $S(x) \geq 0$
- $\exists x \in (0, 1): S(x) = 1$

Damit kommt grundsätzlich jede Abbildung des Einheitsintervalls auf sich selbst, die 0 als Fixpunkt hat, stetig und positiv ist, an mindestens einer Stelle im Intervallinneren die 1 annimmt und im Grenzwert für $x \to 1$ wieder gegen 0 geht, als GSF in Frage. Die zusätzliche Anforderung, die GSF solle dafür sorgen, daß die Vordergrundsequenz gewisse erwünschte Charakteristika aufweist, liefert dann das entscheidende Kriterium für die Eignung einer GSF für ein bestimmtes TES-Modell.

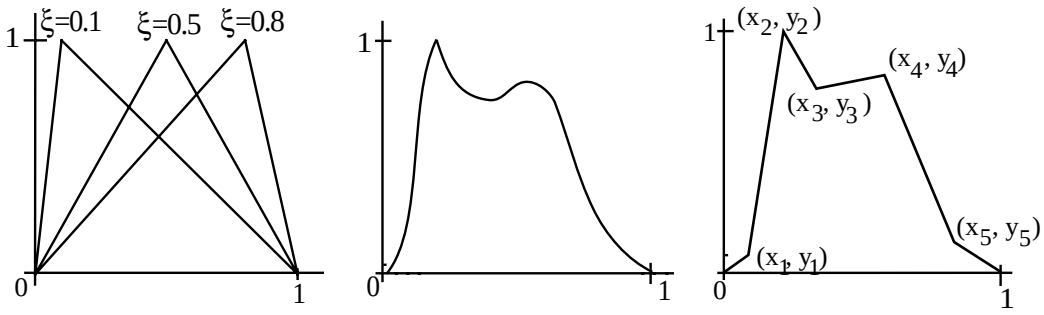


**Abbildung 4-12:** Standard-Stitching-Funktion (links) und zwei Kandidaten für eine Verallgemeinerte Stitching-Funktion (Mitte und rechts)

Berücksichtigt man die in Abschnitt 4.3.1 gemachten Beobachtungen über den Einfluß der SSF auf die entsprechende TES-Vordergrundsequenz, so ergeben sich schnell naheliegende Kandidaten für eine GSF, die zu einer Vordergrundsequenz führt, welche der Referenzkurve von Abbildung 2-3 vom Verlauf her möglichst ähnelt: solche Kandidaten können z.B. der entsprechend auf das Einheitsintervall normierte Durchschnittsverlauf der Referenzkurve während einer Periode sein (Abbildung 4-12 Mitte) oder auch eine geeignet linearisierte Version davon (Abbildung 4-12 rechts).

Wir werden in Abschnitt 4.3.4 sehen, welche Resultate eine derartige Wahl zur Folge hat; in Abschnitt 4.3.6 wird dann schließlich noch eine weiter verfeinerte Variante der GSF angegeben. Zunächst jedoch wird untersucht, welche Konsequenzen die Verallgemeinerung der Stitching-Funktion für die Distortion-Transformation des TES-Schemas hat.

### 4.3.3 Konsequenzen für die Distortion-Transformation

In Bemerkung 4.1 haben wir festgehalten, daß die Hintergrundsequenz auch nach Anwendung der Standard-Stitching-Funktion als Randverteilung die (0,1)-Gleichverteilung aufweist und deshalb die Inversionsmethode nach Lemma A.3 angewendet werden kann. Dies ist im Fall der Verallgemeinerten Stitching-Funktion nicht mehr ohne weiteres der Fall, da die GSF nicht länger mehr eine Gleichverteilung als Randverteilung besitzt. Abbildung 4-13 vergleicht noch einmal eine Standard-Stitching-Funktion mit dem Beispiel einer GSF aus Abschnitt 4.3.2 sowie beide kumulativen Randverteilungen.

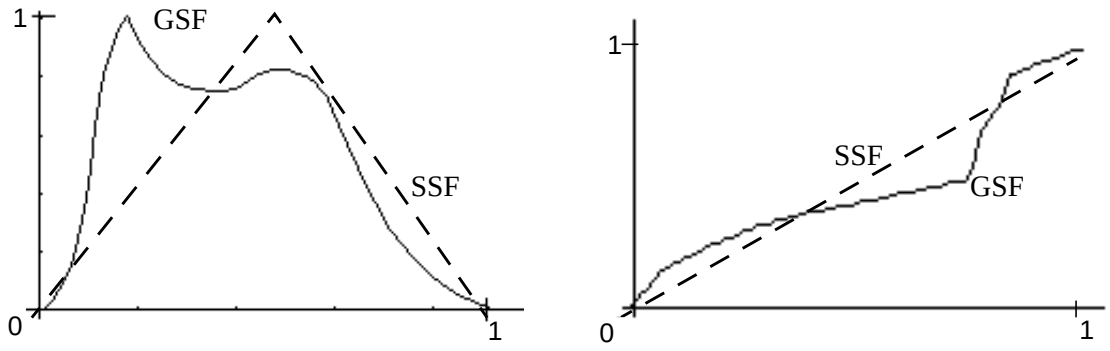


**Abbildung 4-13:** SSF (gestrichelt) und GSF und ihre Randverteilungen

Der Ausweg aus dieser Situation führt über die Anwendung des nachfolgenden Lemmas.

**Lemma 4.3:** *Sei U eine auf dem Intervall* $[0, 1)$ *gleichverteilte Zufallsvariable und S eine Verallgemeinerte Stitching-Funktion mit Randverteilung* $F_S$. *Sei weiterhin F eine beliebige Verteilungsfunktion mit Pseudo-Inverser* $F^{-1}$ *(gem. Definition A.2) sowie*

$$G(x) = F^{-1}(F_S(x)). \tag{4.7}$$

*Die Zufallsvariable*

$$X = G(S(U)) \tag{4.8}$$

*besitzt dann die Verteilungsfunktion F.*

**Beweis ([Rei98a]):**

Sei $F_X(x) = P\{X \le x\}$ mit (4.8) die Verteilungsfunktion der Zufallsvariablen $X$. Mit Definition A.4 ergibt sich

$$F_X(x) = P\{G(S(U)) \le x\} = P\{S(U) \le G^{-1}(x)\} = F_S(G^{-1}(x)). \tag{4.9}$$

Mit (4.7) und der allgemeinen Identität $(a \bullet b)^{-1} = b^{-1} \bullet a^{-1}$ gilt aber

$$G^{-1}(x) = (F^{-1}(F_S(x)))^{-1} = F_S^{-1}(F(x)), \tag{4.10}$$

also mit (4.9)

$$F_X(x) = F_S(G^{-1}(x)) = F_S(F_S^{-1}(F(x))) = F(x). \tag{4.11}$$

q.e.d. ■

Zur Veranschaulichung ist in Abbildung 4-14 (nach [RSH97]) der Übergang von der Standard-TES-Methode zur Verallgemeinerten TES-Methode und die in Lemma 4.3 beschriebenen Konsequenzen noch einmal zusammenfassend skizziert.

Beim Standard-TES (durchgezogene Pfeile) wird die Hintergrundsequenz durch das über die Innovationsverteilung gesteuerte Entlanglaufen an einem Kreis mit Umfang 1 erzeugt (a) und ist gleichverteilt (b). Anwendung der SSF (c) erhält diese Gleichverteilung (d), und die Inversionsmethode ergibt schließlich über die Inverse der vorgegebenen Randverteilungsfunktion (e) die erwünschte Randverteilung (f) (vgl. Abbildung 2-4) der modellierten Sequenz.

Bei der Verallgemeinerten TES-Methode wird auf die gleichverteilte Hintergrundsequenz jetzt eine GSF (g) angewendet, die ihrerseits eine Randverteilung mit Dichte (h) aufweist. In diesem Fall ist nach Lemma 4.3 zunächst die Inverse der Randverteilungsfunktion der GSF (vgl. Abbildung 4-13) zu bestimmen (i), sodann mit (4.7) die Funktion $G(x) = F^{-1}(F_S(x))$ (k), die schließlich – auf die gestitchte Hintergrundsequenz angewendet – wiederum für die resultierende Vordergrundsequenz die korrekten Randverteilung (f) ergibt.

**Bemerkung 4.4:** Alternativ läßt sich der Beweis von Lemma 4.3 auch durch zweimalige Anwendung von Lemma A.3 führen (vgl. [Moh99]). Da $S(U)$ nach Definition A.4 die Verteilung $F_S$ besitzt, ergibt Anwendung des zweiten Teils von Lemma A.3, daß die Zufallsvariable

$$Z = F_S(S(U)) \tag{4.12}$$

eine (0,1)-Gleichverteilung besitzt.

Doch damit erfüllt $Z$ die Voraussetzung des ersten Teils von Lemma A.3, und mit (4.7) kann man folgern, daß

$$F^{-1}(Z) = F^{-1}(F_S(S(U))) = G(S(U)) \tag{4.13}$$

die Verteilung $F$ besitzt, und deshalb mit (4.8) auch die Zufallsvariable $X = G(S(U))$. ■

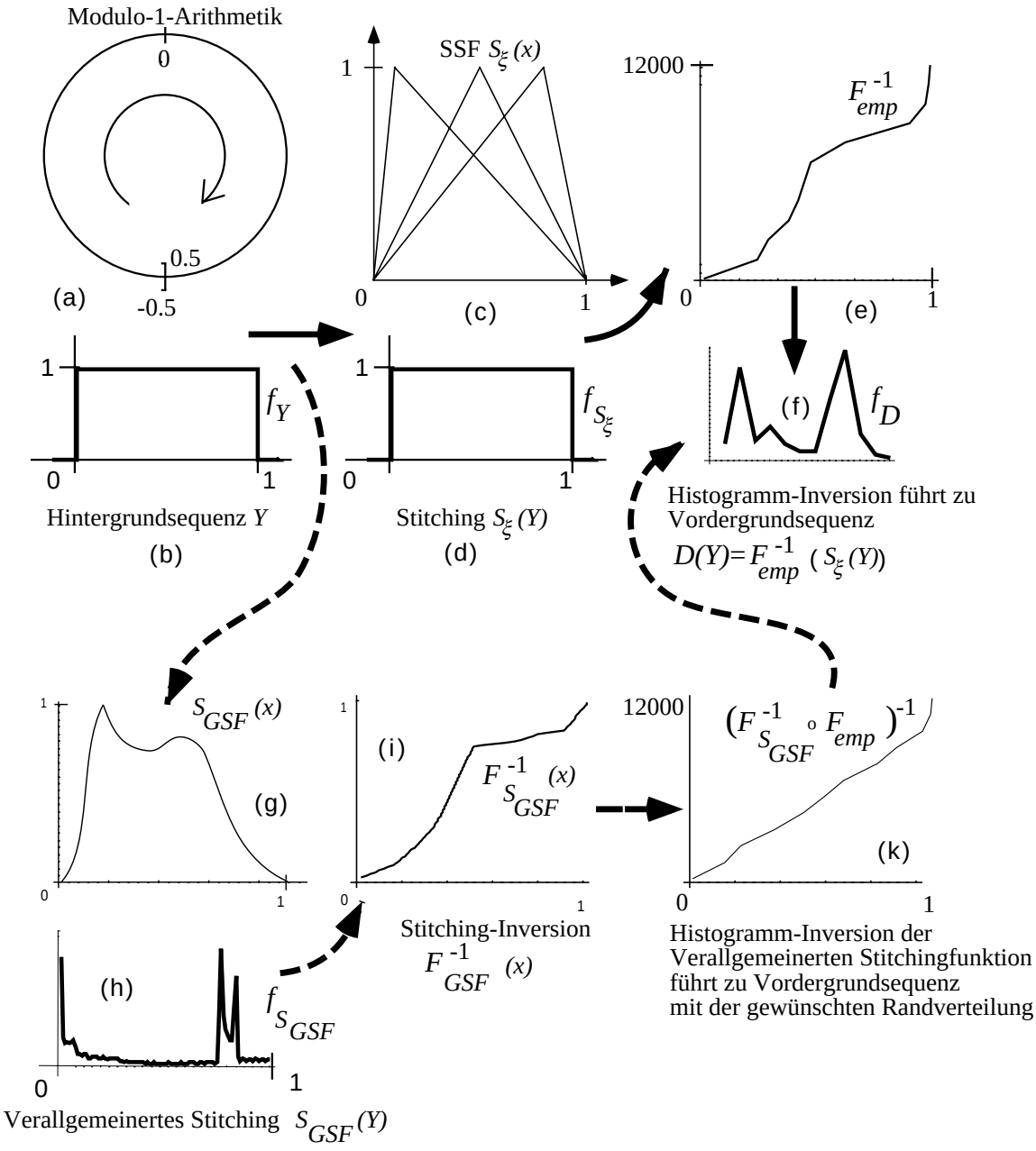


**Abbildung 4-14:** Der Übergang von SSF zu GSF und seine Konsequenzen

Aus dem oben Gesagten ergibt sich unmittelbar eine wichtige Randbedingung für die Wahl der GSF: ihre Randverteilung sollte möglichst effizient berechenbar sein [MRS99]. In diesem Sinne besitzen stückweise lineare GSFs einen enormen Vorteil gegenüber glatteren GSFs

(Abbildung 4-12 Mitte/rechts). Eine stückweise lineare GSF ist dabei durch ihre Stützstellen $(x_i, y_i)$, $i = 0, 1, 2, ..., n$, festgelegt und hat die Form

$$S(x) = \sum_{i=1}^{n} \left( y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}} \cdot (x - x_{i-1}) \right) \cdot 1_{[x_{i-1}, x_i)}(x), \tag{4.14}$$

die sich mittels $m_i = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$ und $b_i = y_{i-1} - m_i x_{i-1}$ zu

$$S(x) = \sum_{i=1}^{n} (m_i x + b_i) \cdot 1_{[x_{i-1}, x_i)}(x) \tag{4.15}$$

vereinfachen läßt.

Lemma 4.5 gibt für stückweise lineare GSFs eine geschlossene Form der Randverteilung an, die eine effiziente Berechnung erlaubt (vgl. [Moh99]).

**Lemma 4.5:** *Sei U eine (0,1)-gleichverteilte Zufallsvariable, S eine stückweise lineare GSF der Form (4.15) mit* $0 = x_0 < x_1 < x_2 < \ldots < x_n = 1$ *und* $y_i = S(x_i)$, $i = 0, 1, 2, ..., n$. *Die Zufallsvariable S(U) besitzt dann folgende Verteilungsfunktion (die Randverteilung der GSF):*

$$\begin{aligned} F_S(y) = & \sum_{i:m_i > 0} \left( \frac{y - b_i}{m_i} - x_{i-1} \right) 1_{[y_{i-1}, y_i)}(y) + (x_i - x_{i-1}) 1_{[y_i, \infty)}(y) \\ & + \sum_{i:m_i < 0} \left( x_i - \frac{y - b_i}{m_i} \right) 1_{[y_i, y_{i-1})}(y) + (x_i - x_{i-1}) 1_{[y_{i-1}, \infty)}(y) \\ & + \sum_{i:m_i = 0} (x_i - x_{i-1}) 1_{[b_i, \infty)}(y) \, . \end{aligned} \tag{4.16}$$

**Beweis:** Für den Beweis werden die einzelnen Monotoniebereiche der GSF (die ja anhand des Vorzeichens der jeweiligen Steigung $m_i$ leicht zu unterscheiden sind) einzeln betrachtet:

$$\begin{aligned} P(S(U) \le y) &= \sum_{i=1}^{n} P((m_i U + b_i) \le y, U \in [x_{i-1}, x_i)) \\ &= \sum_{i:m_i > 0} P\left( U \le \frac{y - b_i}{m_i}, U \in [x_{i-1}, x_i) \right) \\ &\quad + \sum_{i:m_i < 0} P\left( U \ge \frac{y - b_i}{m_i}, U \in [x_{i-1}, x_i) \right) \\ &\quad + \sum_{i:m_i = 0} P(b_i \le y, U \in [x_{i-1}, x_i)). \end{aligned} \tag{4.17}$$

Nun ist für $m_i > 0$

$$P\left(U \le \frac{y-b_i}{m_i}, x_{i-1} \le U < x_i\right) = \begin{cases} \frac{y-b_i}{m_i} - x_{i-1}, & \text{für } m_i x_{i-1} + b_i \le y < m_i x_i + b_i \\ x_i - x_{i-1}, & \text{für } y \ge m_i x_i + b_i \\ 0, & \text{sonst} \end{cases}, \quad (4.18)$$

für $m_i < 0$

$$P\left(U \ge \frac{y-b_i}{m_i}, x_{i-1} \le U < x_i\right) = \begin{cases} x_i - \frac{y-b_i}{m_i}, & \text{für } m_i x_i + b_i \le y < m_i x_{i-1} + b_i \\ x_i - x_{i-1}, & \text{für } y \ge m_i x_{i-1} + b_i \\ 0, & \text{sonst} \end{cases} \quad (4.19)$$

und für $m_i = 0$

$$P(b_i \le y, x_{i-1} \le U < x_i) = \begin{cases} x_i - x_{i-1}, & \text{für } y \ge b_i \\ 0, & \text{sonst} \end{cases} \quad (4.20)$$

Aus der Aufsummierung von (4.18), (4.19) und (4.20) folgt die Behauptung. ■

Lemma 4.5 wird im Anhang A.1 anhand der exemplarischen Berechnung von Beispiel A.5 noch veranschaulicht.

### 4.3.4 Ergebnisse für die Referenzkurve

Aufbauend auf den Resultaten der vorhergehenden Abschnitte wurde das Modellierungstool TESter zur Modellierung von periodischem Verkehr mit der Verallgemeinerten TES-Methode entwickelt (vgl. [Rei98b]) und in [Moh99] optimiert. Nachfolgend werden einige der damit erzielten Ergebnisse dargestellt.

Abbildung 4-15 zeigt im Vergleich zu Abbildung 4-9, daß sich durch Einführung der Verallgemeinerten Stitching-Funktion (in diesem Beispiel gemäß Abbildung 4-12 Mitte) die Erfüllung des Kriteriums (iii) der visuellen Ähnlichkeit signifikant verbessert hat. Das Mittagsloch ist klar zu erkennen, die Periodizität wird im wesentlichen eingehalten, und auch die Autokorrelationsfunktionen stimmen in den führenden Lags gut überein.

Die Vermutung liegt nahe, daß die Ursache sowohl für die Störung in der Periodizität, wie sie in Abbildung 4-15 Mitte vor allem um den x-Wert 80 ersichtlich ist, als auch für die Abweichung der Autokorrelationsfunktionen in der Wahl eines relativ breiten Trägerintervalls der Innovationsdichte ($\delta = 0.05$, Abbildung 4-15 rechts) zu suchen ist. Deshalb wurde der Einfluß dieser Breite $\delta$ als nächstes untersucht.

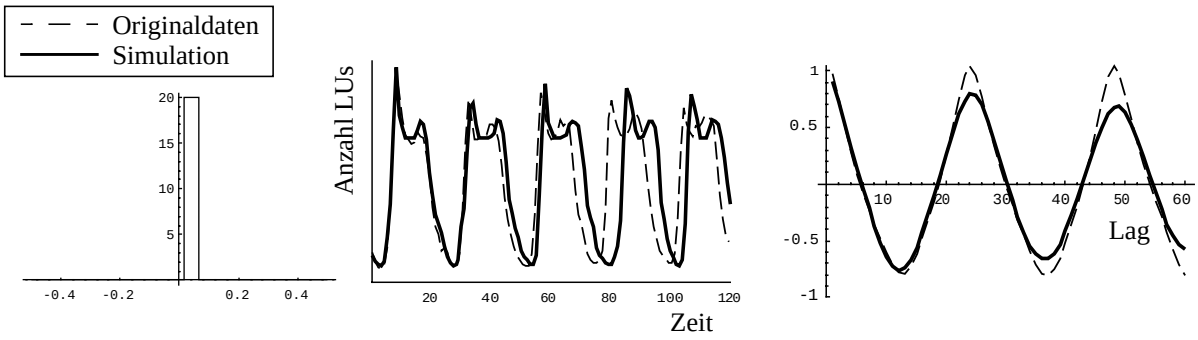


**Abbildung 4-15:** GTES-Modell der Referenzkurve

Abbildung 4-16 zeigt in der oberen Reihe das Resultat für einen sehr schmalen Träger der Innovationsdichte ($\delta = 0.01$) und im Vergleich dazu das Ergebnis für einen ziemlich breiten Träger ($\delta = 0.1$). Im Vergleich zu Abbildung 4-15 finden wir unsere Vermutung bestätigt: Der sehr schmale Träger sorgt für eine nahezu deterministische Hintergrundsequenz, demzufolge weist die Vordergrundsequenz kaum Unterschiede von einer Periode zur anderen auf, dafür ist die Periodizität sehr gut erfüllt und die Autokorrelationsfunktion der modellierten Sequenz ist nahezu identisch zur vorgegebenen der Referenzmessung. Im Gegensatz dazu sorgt eine weitere Verbreiterung des Trägers zu einer unübersehbaren Verschlechterung hinsichtlich der Autokorrelationen, und die Vordergrundsequenz läuft sehr bald gänzlich aus der Periode, was allerdings einen deutlich lebhafteren Kurvenverlauf mit sich bringt.

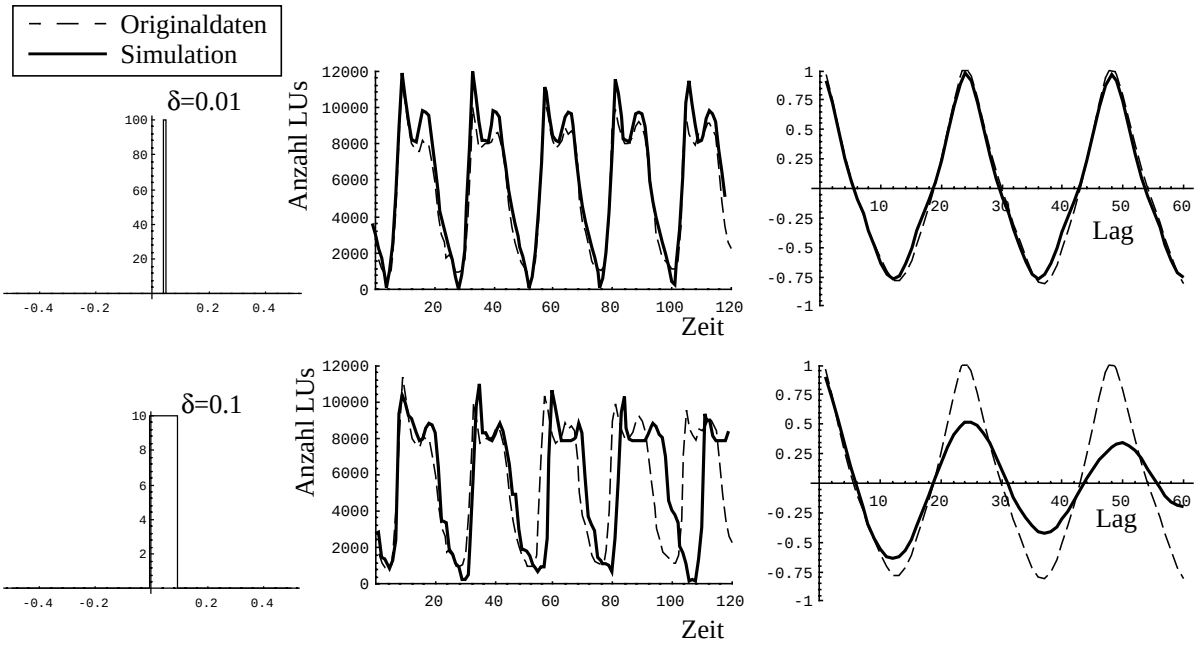


**Abbildung 4-16:** GTES-Modell der Referenzkurve: Einfluß der Trägerbreite der Innovationsdichte

Weiterhin wurde untersucht, ob sich das Ergebnis durch Verwendung einer aus zwei separaten Teilintervallen bestehenden Innovationsdichte verbessern läßt. Insbesondere wurden die Träger beider Intervalle zunächst gleich breit gewählt, wobei ein Trägerintervall symmetrisch zum Ursprung lag. Dieser sogenannte “Progress-Retard Approach”[5] [RSH97] führte zu den in Abbildung 4-17 dargestellten Ergebnissen.

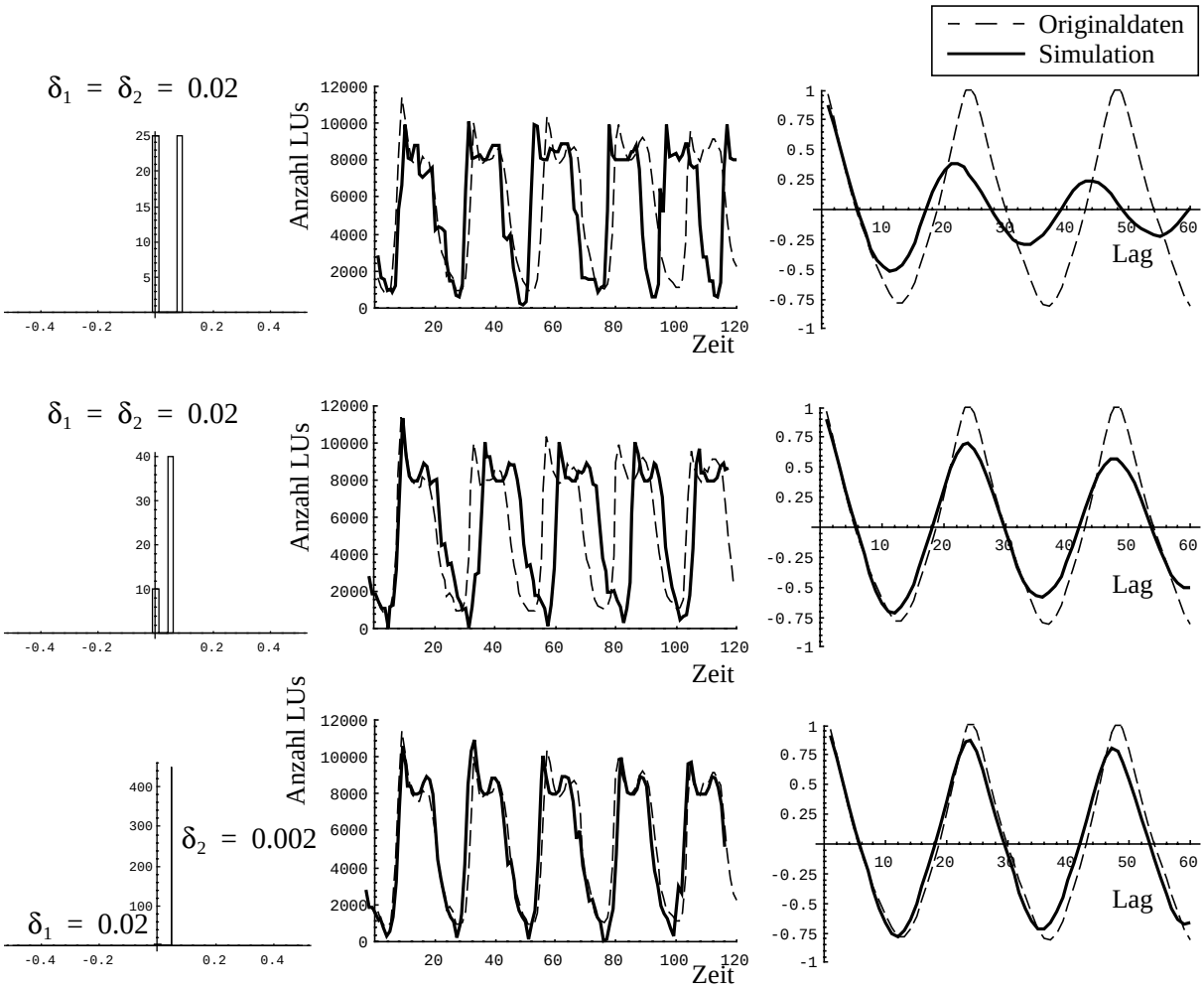


**Abbildung 4-17:** Progress-Retard-Modelle für die Referenzkurve

In den ersten beiden Reihen von Abbildung 4-17 sind die beiden Trägerintervalle jeweils $\delta = 0.02$ breit, im ersten Fall zudem gleichmäßig gewichtet, was zu deutlichen Periodenverschiebungen in beide Richtungen sowie schlechter Übereinstimmung der Autokorrelationsfunktionen führt. Daraufhin wurde für die mittlere Reihe die Gewichtung zwischen beiden Intervallen zu einem Verhältnis von 20% zu 80% verändert; dies hatte wohltuenden Einfluß auf die Autokorrelationsfunktionen, die Periodizität ging zumindest nicht gänzlich verloren, und die Vordergrundsequenzen weisen keineswegs einen starren Verlauf auf.

5. Die genannte Form der Innovationsdichte hat zur Folge, daß der Verlauf der Hintergrundsequenz geprägt ist durch die Abwechslung von größeren Sprüngen mit eher marginalem “Vorwärtskriechen”, was den gewählten Namen erklären mag.

In einem dritten Modell (Abbildung 4-17 untere Reihe) wurde schließlich eine Extremsituation erkundet: Diesmal hatte das linke (im Ursprung zentrierte) Intervall immer noch eine Breite von 0.02, während für das andere Intervall die Breite 0.002 gewählt wurde. Zudem wurde die Gewichtung zwischen beiden auf 10% zu 90% festgelegt. Diesmal zeigt sich ein zufriedenstellenderes Ergebnis: Die Vordergrundsequenz ist hinreichend periodisch, hat aber immer noch von Periode zu Periode hinreichend große zufällige Unterschiede, und die Autokorrelationsfunktionen stimmen recht gut überein, so daß man dieser Parametrisierung eine für den untersuchten Zweck ausreichende Qualität zusprechen kann.

### 4.3.5 Weitere Anwendung: ein Generalized TES-Modell zur Modellierung von MPEG-codiertem Videoverkehr

Als weiteres Beispiel für die Anwendung eines Verallgemeinerten TES-Modells wurde die Modellierung von Videoverkehr, der mittels MPEG codiert wurde, untersucht. MPEG (Moving Pictures Expert Group) ist ein Standard für die Bildcodierung [LeG91], der unter den folgenden zwei Gesichtspunkten entwickelt wurde: zum einen sollte er eine effiziente Codierung gewährleisten, zum anderen aber auch eine schnellen Zugriff (Random Access) auf willkürliche Bilder garantieren. Deshalb erweist er sich bei näherem Hinschauen als Kompromiß zwischen Interframe- und Intraframe-Codierung: erstere nutzt hierbei die temporale Redundanz zwischen aufeinanderfolgenden Bildern eines Films aus, um die Datenmenge zu reduzieren, während letztere den willkürlichen Bildzugriff ermöglicht sowie von Zeit zu Zeit eine Auffrischung der Bilder erlaubt. Zu diesem Zweck definiert der MPEG-Standard [iso93] vier verschiedene Typen von Bildcodierung: I-Frames (Intra-coded Frames), P-Frames (Predictive-coded Frames), B-Frames (Bi-directionally coded Frames) und D-Frames (DC-coded Frames).

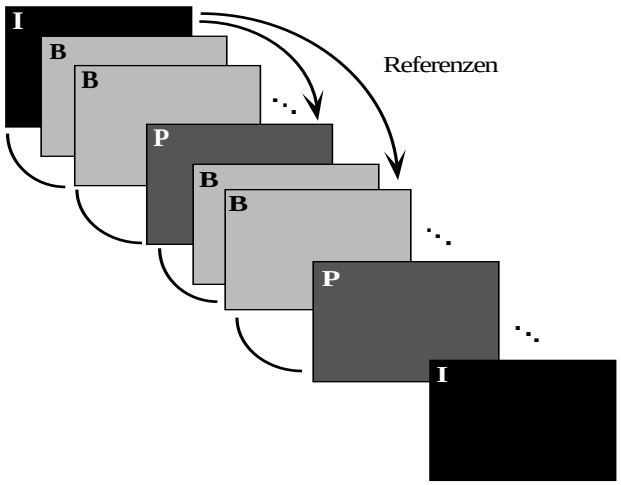


**Abbildung 4-18:** MPEG-Frames und ihre Abhängigkeit untereinander

Abbildung 4-18 zeigt die Abhängigkeit der einzelnen Frametypen untereinander: Während I-Frames ohne Referenz auf andere Rahmen codiert sind, d. h. jeweils die komplette Information eines gesamten Bildes enthalten, und somit geeignete Punkte für den willkürlichen Zugriff innerhalb eines MPEG-Stroms darstellen, benötigen P-Frames Informationen des vorausgehenden I-Frames und/oder der vorhergehenden anderen P-Frames. Sie basieren auf zeitlicher Redundanz, da aufeinanderfolgende Bilder in einem Videostrom oft Regionen aufweisen, die zwar räumlich verschoben, aber ansonsten unverändert geblieben sind. Daher genügt es prinzipiell, den Bewegungsvektor einer solchen Region sowie die (relativ geringfügige) Differenz, die von Änderungen innerhalb der Region selbst herrührt, zu codieren. Es hat sich herausgestellt, daß dieses Prinzip beträchtliche Einsparungen in der Datenmenge erlaubt. B-Frames gehen schließlich noch ein Stück weiter, da sie Informationen sowohl von vorhergehenden als auch darauffolgenden I-Frames und P-Frames benötigen. In einem B-Frame ist somit die Differenz zwischen dem aktuellen Rahmen und einer Art Interpolation aus vergangenen und zukünftigen Rahmen codiert, was letzten Endes eine enorm hohe Kompressionsrate erlaubt. D-Frames schließlich lassen sich als ein Sonderfall der I-Frames auffassen und spielen in der Praxis meist keine große Rolle.

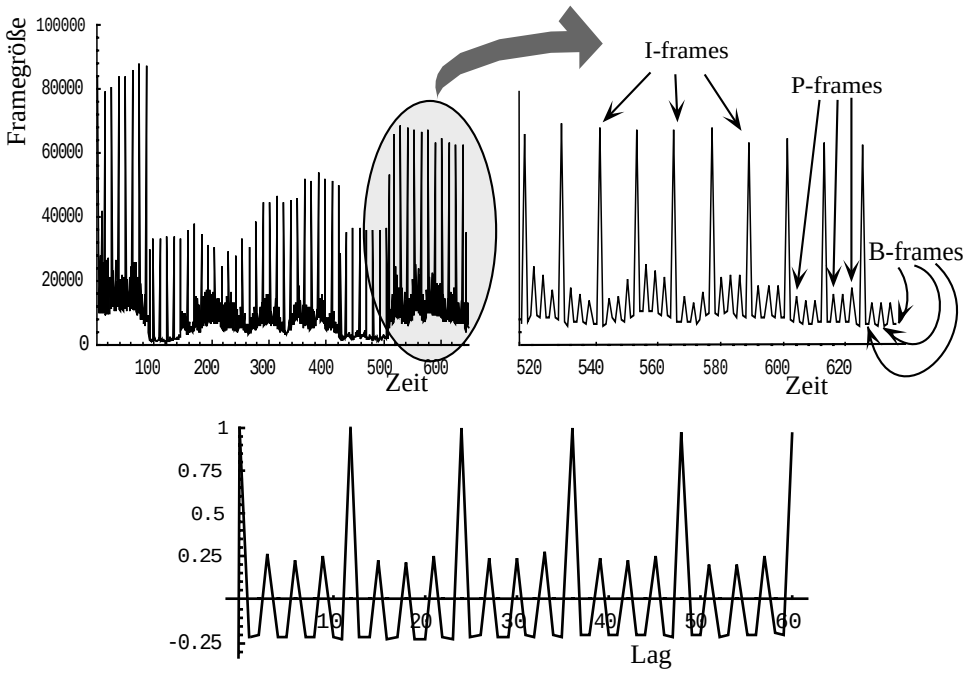


**Abbildung 4-19:** Beispiel für eine Rahmensequenz am Beginn des Films “Das Schweigen der Lämmer” [Ro95] (oben links, oben rechts ein vergrößerter Teilabschnitt, nach rechts aufgetragen jeweils die Rahmennummer) sowie zugehörige Autokorrelationsfunktion (unten)

Die Abfolge der Sequenz von I-, P- und B-Frames wird von der entsprechenden MPEG-Applikation bestimmt. Von größerer Beliebtheit ist beispielsweise die Sequenz

IBBPBBPBBPBB IBBPBBPBBPBB IBBPBBPBBPBB..., (4.21)

die eine Zugriffsauflösung von ca. 330 msec. erlaubt. Auf diese Weise ist etwa das in Abbildung 4-19 gezeigte Beispiel codiert, das den ersten 20 Sekunden des Films “Das Schweigen der Lämmer” entnommen ist [Ro95]. In dieser Abbildung wurden hierbei die einzelnen diskreten Datenpunkte, die den Größen einzelner Rahmen entsprechen, linear verbunden; daher entsprechen die Minima jeweils einzelnen B-Frames und nicht - wie in anderen Abbildungen oft üblich - irgendwelchen “idle periods”.

Abbildung 4-19 links oben zeigt uns das Auftreten einzelner Muster, die jeweils verschiedenen Kameraeinstellungen entsprechen (da die Übertragungsfrequenz 29.97 Rahmen pro Sekunde beträgt, entsprechen 100 Rahmen etwa 3 Filmsekunden). Innerhalb einer Kameraeinstellung (wie sie z. B. in derselben Abbildung oben rechts herausgezoomt wurde), ist der Verlauf der einzelnen Rahmengrößen dagegen ziemlich regelmäßig. In unserem Beispiel haben die I-Frames hierbei jeweils etwa die zehnfache Größe der B-Frames, die P-Frames liegen jeweils dazwischen. Die Verwendung des Schemas 4.21 führt zu der strikten Periode von 12, die sich auch in der Autokorrelationsfunktion (Abbildung 4-19 unten) deutlich widerspiegelt.

In der Literatur gibt es seit langem Ansätze zur Modellierung von MPEG-Strömen mit Hilfe der TES-Methode, z. B. [MS93], [LMRS94], [RMR94] oder [ILDK95][6]. Hierbei wird aber in aller Regel für jeden individuellen Rahmentyp ein eigenes TES-Modell erstellt. Im Gegensatz hierzu lassen sich mit unserem Ansatz [Rei98a] alle drei Rahmentypen innerhalb eines einzigen Modells zusammenführen, während der Fokus nach wie vor auf der Modellierung einzelner Kameraeinstellungen (Takes) liegt.

Für die Modellerstellung ist zu beachten, daß trotz der hohen zu beobachtenden Regularität die einzelnen Takes sich doch voneinander insofern unterscheiden, als die Frames innerhalb eines Takes (verglichen mit anderen) gewöhnlich relativ einheitlich groß sind. Um dem Rechnung zu tragen, wurde eine weiter verfeinerte Variante der Generalized Stitching Function entwickelt.

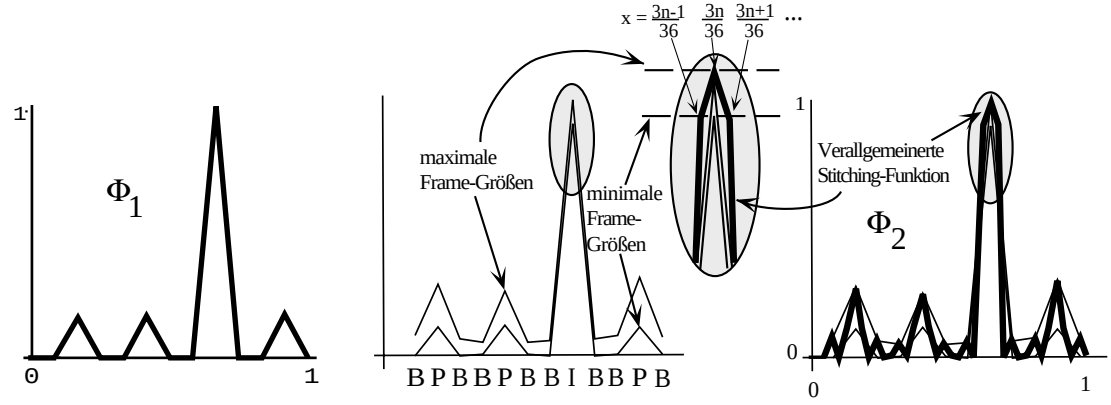


**Abbildung 4-20:** Zwei Varianten einer GSF für das MPEG-Beispiel

6. Eine allgemeinere Behandlung von Modellierungsmethoden für MPEG findet sich in [Ro97].

Abbildung 4-20 links zeigt die auf dem normalen Weg über Durchschnittsbildung gewonnene GSF $\Phi_1$, während in der Mitte jeweils die maximalen und minimalen Rahmengrößen $\Phi^{max}$ bzw. $\Phi^{min}$ jeweils für die einzelnen Rahmen der Sequenz 4.21 dargestellt sind. Aus dieser zusätzlichen Information läßt sich die verfeinerte Variante $\Phi_2$ folgendermaßen gewinnen: Definiere für $n = 1, 2, 3, \ldots, 12$

$$\Phi_2(x_n) = \begin{cases} \Phi^{min}(x_n) & \text{falls } \left(x_n = \frac{3n-1}{36}\right) \vee \left(x_n = \frac{3n+1}{36}\right) mod 1 \\ \Phi^{max}(x_n) & \text{falls } x_n = \frac{3n}{36} \end{cases} \tag{4.22}$$

Die verfeinerte GSF wie in Abbildung 4-20 rechts dargestellt ergibt sich dann einfach durch lineare Interpolation.

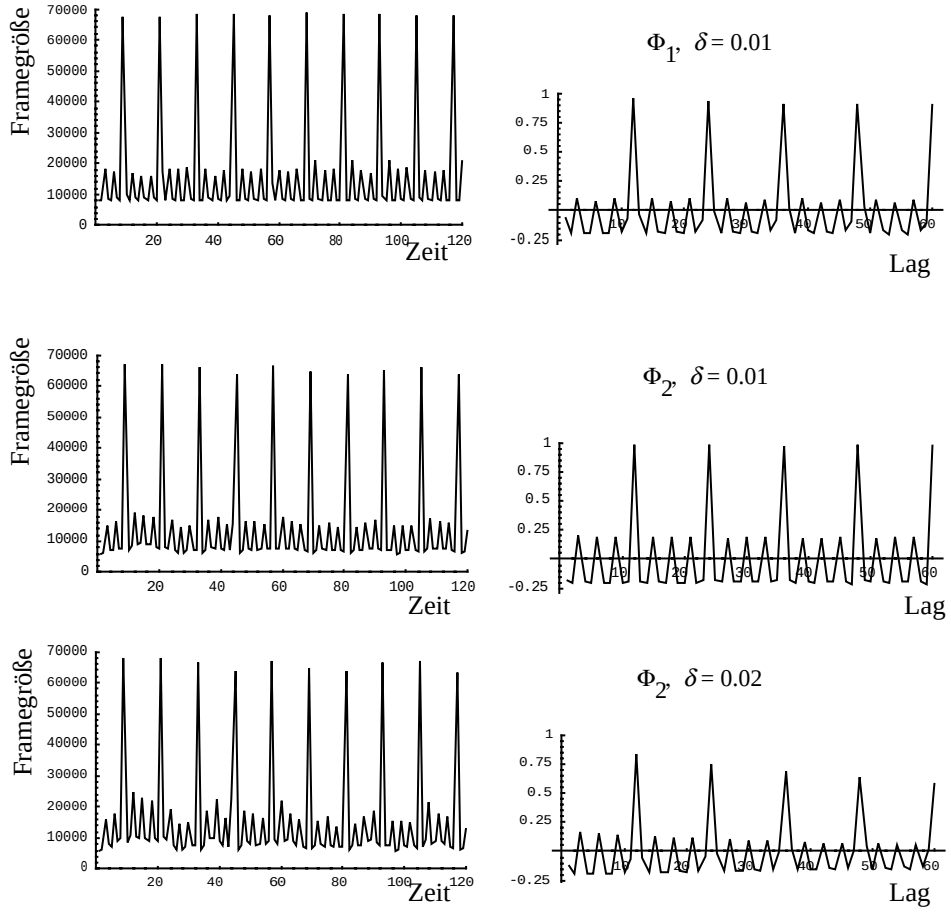


**Abbildung 4-21:** Vordergrundsequenz des GTES-Modells für verschiedene Innovationsdichten und verschiedene GSF-Varianten sowie zugehörige Autokorrelationsfunktionen

Durch dieses Vorgehen kann man erreichen, daß bei Verwendung einer Innovationsdichte mit relativ kleinem Trägerintervall (z. B. $\delta = 0.2$) die Hintergrundsequenz sich jeweils in dem zu

einem bestimmten Rahmen gehörenden Intervall bewegt und daher die Rahmengröße, wie von der Vordergrundsequenz angegeben, zwischen dem jeweiligen Minimum und Maximum liegt.

Abbildung 4-21 zeigt exemplarisch den Einfluß von GSF und Innovationsdichte auf das Modellierungsergebnis. Die Periodizität wurde hierbei durch die Verwendung einer gleichverteilten Innovationsdichte mit einem kleinen Träger, nämlich $\left[\frac{1}{12}-\frac{\delta}{2};\frac{1}{12}+\frac{\delta}{2}\right]$, erreicht. In der linken Spalte sind die Vordergrundsequenzen, rechts die zugehörigen Autokorrelationsfunktionen dargestellt. Im Vergleich der ersten beiden Reihen wird deutlich, daß der Übergang von $\Phi_1$ zu $\Phi_2$ eine erkennbare Verbesserung der Modellierung bewirkt: Die Verwendung der einfachen GSF führt trotz relativ deterministischer Hintergrundsequenz zu größeren Abweichungen hinsichtlich Periodizität und Autokorrelationsfunktion, während die verfeinerte Variante von Reihe 2 eine Vordergrundsequenz erzielt, die kaum vom Original zu unterscheiden ist und obendrein noch strikt die Periodizität erhält. Die Verdoppelung des Trägerintervalls in der dritten Reihe führt wie erwartet zu einem lebhafteren Kurvenverlauf.

### 4.3.6 Die verfeinerte GSF im Referenzproblem

Die in Abschnitt 4.3.5 vorgeschlagene Verfeinerung der Generalized Stitching Function läßt sich natürlich auch im Referenzfall der Modellierung von GSM Location Updates einsetzen [Rei98b], wie er weiter oben schon ausführlich behandelt worden ist. Das Vorgehen bleibt dabei identisch: Zunächst werden $\Phi^{max}$ und $\Phi^{min}$ definiert, die für jede Stunde den jeweiligen Maximal- bzw. Minimalwert aus der gemessenen Zeitreihe von Abbildung 2-3 beschreibt (d.h. das Maximum bzw. Minimum aus den gemessenen 5 Samples pro Uhrzeit). Dann erhält man die verfeinerte GSF als lineare Interpolation zu den Stützstellen ($n = 1, 2, \ldots, 24$)

$$\Phi_2(x_n) = \begin{cases} \Phi^{min}(x_n) & \text{falls} \quad \left(x_n = \frac{3n-1}{72}\right) \vee \left(x_n = \frac{3n+1}{72}\right) mod\, 1 \\ \Phi^{max}(x_n) & \text{falls} \quad x_n = \frac{3n}{72} \end{cases} . \qquad (4.23)$$

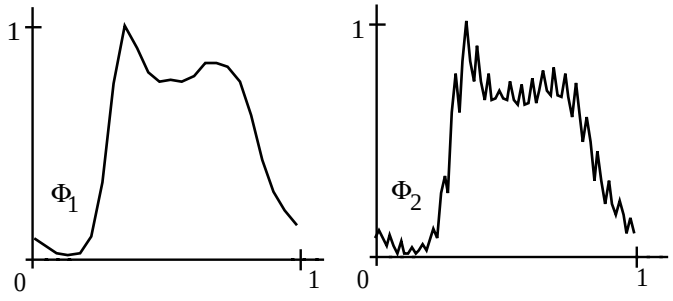


**Abbildung 4-22:** Beide Varianten der GSF im Referenzbeispiel

Abbildung 4-22 zeigt beide Varianten der GSF. In Abbildung 4-23 wird exemplarisch das Simulationsergebnis für die so gewonnenen zwei GSF-Varianten und zwei unterschiedliche Innovationsdichten gezeigt. Wiederum rührt die Periodizität von der Verwendung von gleichverteilten Innovationsdichten, diesmal der Form $\left[\frac{1}{24}-\frac{\delta}{2};\frac{1}{24}+\frac{\delta}{2}\right]$, her.

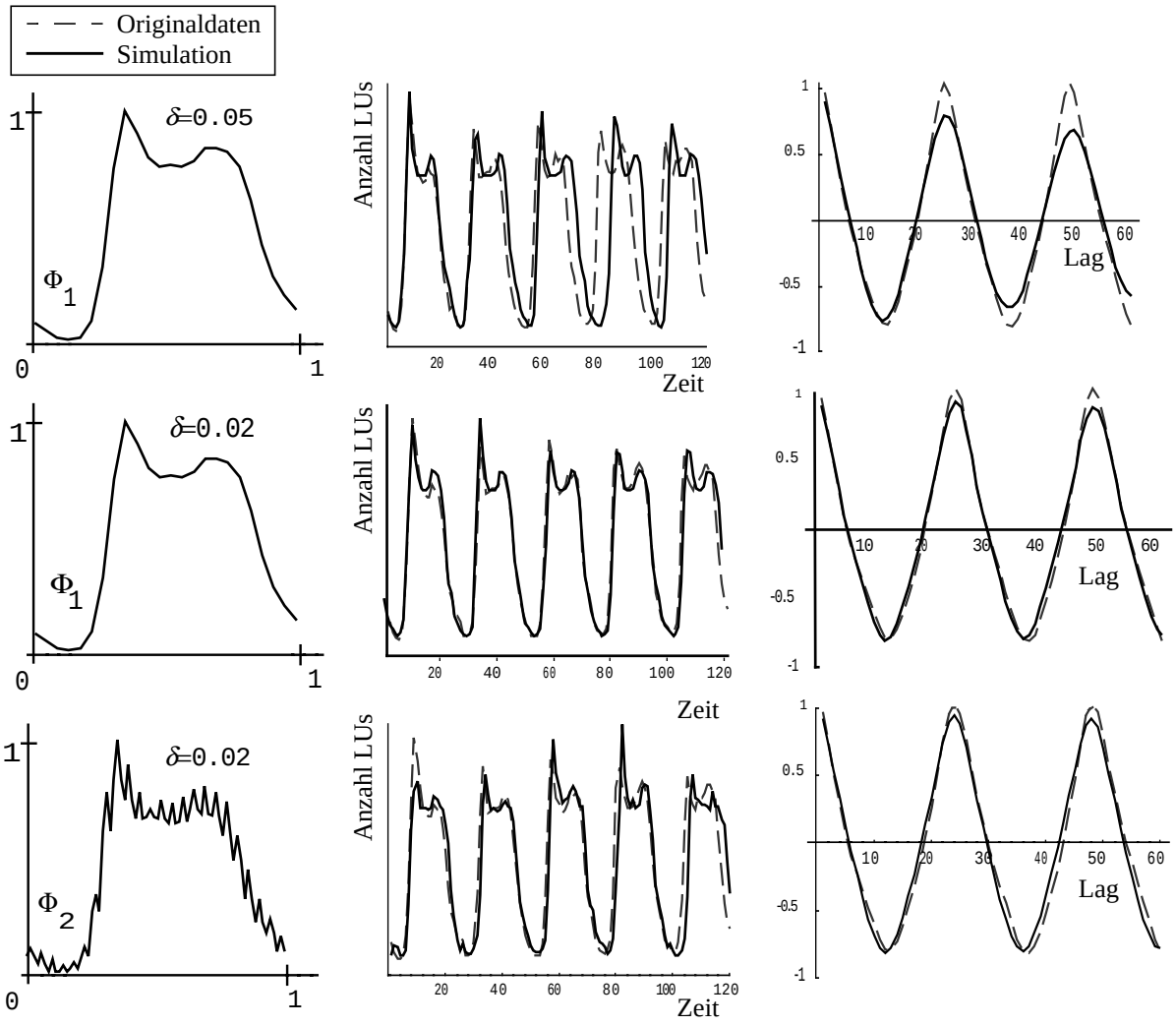


**Abbildung 4-23:** Simulationsergebnisse für verschiedene δ und beide GSF-Varianten. Simulationsergebnisse und Autokorrelationsfunktionen (Referenzreihe grau gestrichelt)

Man kann beobachten, daß bei kleinem Träger der Innovationsdichte die führenden Autokorrelationen gut nachgebildet werden (mittlere und untere Reihe rechts). Im Fall der einfachen GSF (mittlere Reihe Mitte) ist die Vordergrundsequenz jedoch deutlich deterministischer als bei der verfeinerten GSF (untere Reihe Mitte). Ein ähnlich lebhafter Kurvenverlauf läßt sich durch Verbreiterung des Trägers auch bei einfacher GSF erreichen (obere Reihe Mitte), wird jedoch durch ein deutlich schwächeres Resultat bei der Autokorrelationsfunktion (obere Reihe rechts) erkauft, so daß letztendlich dem Modell der verfeinerten GSF bei schmalem Träger der Innovationsdichte der Vorzug gebührt.

# 4.4 Automatisierung von TES

## 4.4.1 Grundsätzliche Überlegungen und Vorgehensweise

In den vorangegangenen Abschnitten hat sich immer wieder die Wahl einer geeigneten Innovationsdichte als entscheidend für die Qualität eines TES-Modells herausgestellt. Die hierfür erforderliche Suche über einen großen Parameterraum (selbst bei Beschränkung auf Innovationsdichten der Form (4.6) wird jedes Modell mindestens durch die Dimensionen Intervallzahl, Breite der jeweiligen Träger sowie gegenseitige Gewichtung charakterisiert) kann entweder – wie im bisherigen Verlauf praktiziert – heuristisch durchgeführt oder automatisiert werden. In der Literatur findet sich bislang lediglich ein Versuch, die Parameter eines (nicht verallgemeinerten) TES-Modells automatisch zu bestimmen [JM95]. Es ist instruktiv, an dieser Stelle einen kurzen Blick auf den dabei verwendeten Ansatz zu werfen.

Die Grundidee des Algorithmus beruht auf der Minimierung des Abstandes zwischen der Autokorrelationsfunktion der Modelldaten und ihrem empirischen Gegenstück. Die zugrundeliegende Metrik basiert dabei auf einer gewichteten Summe der quadrierten Differenzen zwischen den Autokorrelationen entsprechender Lags; die zu minimierende Zielfunktion hat also die Form

$$g(f_V, \xi) = \sum_{\tau=1}^{T} a_\tau (R_{\text{emp}}(\tau) - R_{\text{mod}}(\tau))^2 \tag{4.24}$$

wobei $R_{\text{emp}}(\tau)$ die vorgegebene empirische Autokorrelation zum Lag $\tau$ bezeichnet und $R_{\text{mod}}(\tau)$ das entsprechende Gegenstück aus dem TES-Modell; in die Gewichtung $a_\tau$ geht vor allem die Tatsache ein, daß für mehr oder weniger alle Anwendungen die Übereinstimmung der Autokorrelationen für kleine Lags bedeutsamer ist als für große. Optimiert wird über die Innovationsdichte $f_V$ und den Stitching-Parameter $\xi$, wie sie aus Abschnitt 4.1 bekannt sind.

Entscheidend ist nun, daß gemäß (4.3) eine analytische Darstellung der Funktion $R_{\text{mod}}(\tau)$ besteht, die eine schnelle und numerisch stabile Berechnung der Zielfunktion und ihrer partiellen Ableitungen erlaubt, auch wenn die resultierenden Formeln äußerst länglich geraten (s. [JM95]), weshalb auf ihre detaillierte Wiedergabe hier verzichtet werden soll. Unter Verwendung des GSLO-Algorithmus' (Global Search Local Optimization), der lokale nichtlineare Optimierung mit einer globalen Suche nach dem Minimum der Zielfunktion verbindet, lassen sich auf diese Weise Parametrisierungen in Form von Paaren $(f_V, \xi)$ angeben, die die Zielfunktion (4.24) minimieren.

Erste Implementationsversuche im Rahmen von Diplomarbeiten ([Do97], [E-H97]) haben ergeben, daß dieser Ansatz aufgrund des Umfangs der zugrundeliegenden Berechnungen nur unter großem Aufwand realisierbar erscheint. Da außerdem die Einbeziehung Verallgemeinerter Stitching-Funktionen die Komplexität dieses Algorithmus' (insbesondere durch die Erwei-

terung des Laplace-transformierten Verzerrungsterms in (4.3)) nochmals signifikant erhöht, wurde ein neues Automatisierungsverfahren entwickelt, das speziell zur Modellierung von Zeitreihen mit periodischem Verhalten und mit Hilfe des Konzepts der GSF geeignet ist und sich als wesentlich einfacher als das in [JM95] vorgeschlagene Verfahren erweist. Um den Parameterraum von vorneherein hinreichend einzuschränken, ist die Suche nach einer geeigneten Innovationsverteilung bei diesem Verfahren auf Verteilungsfamilien mit sehr wenigen Parametern beschränkt. Als besonders geeignet hat sich die Familie der Gammaverteilungen (vgl. Anhang A.1 Definition A.6) erwiesen, weil sich mit ihrer Hilfe alle Variationskoeffizienten $c_X > 0$ abdecken lassen (vgl. [Bau91]) und eine Gammaverteilung sich bereits durch zwei Parameter, die man im wesentlichen aus Erwartungswert und Varianz berechnen kann, eindeutig charakterisieren läßt (Näheres hierzu im Anhang A.1).

### 4.4.2 Der Automatisierungsalgorithmus

Aus der im vorigen Abschnitt beschriebenen Grundidee heraus läßt sich folgender Algorithmus zur Bestimmung einer geeigneten Innovationsverteilung formulieren:

1. *Festlegung eines Histogramm der empirischen Zeitreihe.*

   Im ersten Schritt wird eine Einteilung der vorgegebenen Meßwerte in für eine Histogrammerstellung geeignete Klassen vorgenommen. Die untere Grenze der ersten Klasse wird durch $\alpha(F) = sup\{x|F(x) = 0\}$, die obere Grenze der letzten Klasse durch $\omega(F) = inf\{x|F(x) = 1\}$ gebildet, wobei $F$ die vorgegebene Randverteilung der Referenzsequenz bezeichnet.

2. *Auswahl einer typische Periode mit n Werten der Meßreihe; der linke Rand der Verteilung wird am Anfang und am Ende dieser Periode hinzugefügt.*

   Setze also $x_0 = x_{n+1} = \alpha(F)$ und wähle zu $x_1 < x_2 < \ldots x_n$ zugehörige Meßwerte $y_1, \ldots, y_n$ einer typischen Periode. Bisher haben wir stets den Durchschnitt aller Perioden der Referenzkurve als “typische Periode” verwendet, was auch diesen Schritt automatisch ablaufen zu lassen erlaubt. Es ist aber auch denkbar, die Auswahl einer typischen Periode von Hand zuzulassen.

3. *Approximation der Meßwerte der typischen Periode durch eine stückweise lineare Funktion. Durch Skalierung auf das Intervall* $[0, 1]$ *entsteht daraus eine Stitching-Funktion.*

   Hierzu werden zunächst die Monotonieintervalle der typischen Periode bestimmt: Suche die eindeutig bestimmten $n_1 < \ldots < n_k \in \{1, \ldots, n\}$, für die gilt:

   - $y_1 < \ldots < y_{n_1}$ mit $y_{n_1} > y_{n_1+1}$;
   - $y_{n_1} > \ldots > y_{n_2}$ mit $y_{n_2} < y_{n_2+1}$;
   - …

Wähle sodann für $i = 1, \ldots, k$

$$m_i \in \left( \frac{n_i}{n+1}, \frac{n_i+1}{n+1} \right). \tag{4.25}$$

Bestimme schließlich eine stückweise lineare Stitching-Funktion $S$, die den folgenden Bedingungen genügt:

- $S$ streng monoton steigend in $[0, m_1]$ mit $S(m_1) \geq \dfrac{x_{n_1}}{\omega(F) - \alpha(F)}$
- $S$ streng monoton fallend in $[m_1, m_2]$ mit $S(m_1) \leq \dfrac{x_{n_2}}{\omega(F) - \alpha(F)}$
- …
- $S$ streng monoton fallend in $[m_k, 1]$.

4. *Berechnung der Randverteilungsfunktion $F_S$ der Verallgemeinerten Stitching-Funktion.*

   Hierzu wird gemäß Lemma 4.5 vorgegangen. Eine Beispielrechnung findet sich im Anhang A.1 als Beispiel A.5.

5. *Erzeugung gleichverteilter Zufallsvariablen auf dem Intervall $[0, 1)$ mit Hilfe der inversen Verteilungsfunktion.*

   Berechne $z_i = F_S^{-1}(F(y_i))$. Nach Lemma A.3 ergibt dies eine (0,1)-gleichverteilte Zufallsvariable. Da $F_S$ streng monoton steigend ist, ist auch $F_S^{-1}$ isoton. Deshalb besitzt die Sequenz $(z_i)_{i = 1, 2, \ldots, n}$ dieselben Monotonieeigenschaften wie die vorgegebene Meßreihe.

6. *Aufteilung der Stitching-Funktion in monotone Teilintervalle.*

   Setze $S_1 = S|_{[0, m_1]}$, $S_2 = S|_{[m_1, m_2]}$, ..., $S_{k+1} = S|_{[m_k, 1]}$
   Dadurch wird die Stitching-Funktion intervallweise invertierbar.

7. *Transformation der Teilintervalle durch die jeweiligen Inversen.*

   Bestimme die Hintergrundsequenz $u_1, \ldots, u_n$:

   - $u_i = \begin{cases} S_1^{-1}(z_i) \Leftrightarrow z_i < S(m_1) \\ m_1 \qquad \text{sonst} \end{cases}$ für $i = 1, 2, \ldots, n_1$
   - $u_i = \begin{cases} S_2^{-1}(z_i) \Leftrightarrow z_i < S(m_2) \\ m_2 \qquad \text{sonst} \end{cases}$ für $i = n_1 + 1, \ldots, n_2$
   - …

- $u_i = S_{k+1}^{-1}(z_i)$ für $i = n_k + 1, \ldots, n$.

  Hier werden also die Daten der Teilintervalle durch die jeweiligen Inversen transformiert. Bei ungünstiger Wahl der Stitching-Funktion kann es vorkommen, daß einige $z_i$ außerhalb des Definitionsbereiches liegen; daher wird in diesem Fall der Endpunkt des Definitionsbereichs eingesetzt.

*8. Schätzung des Erwartungswertes und der Varianz der als zugrundeliegend angenommenen Gammaverteilung mit Hilfe des Standardverfahrens der schließenden Statistik.*

Setze für $i = 1, \ldots, n-1$

$$v_i = u_{i+1} - u_i \tag{4.26}$$

und berechne damit

$$\tilde{\mu}_V = \frac{1}{n-1} \sum_{i=1}^{n-1} v_i \tag{4.27}$$

und

$$\tilde{\sigma}_V^2 = \frac{1}{n-2} \sum_{i=1}^{n-1} (v_i - \tilde{\mu}_V)^2 . \tag{4.28}$$

*9. Berechnung der Parameter der Gammaverteilung gemäß Anhang A.1.*

Berechne schließlich

$$\alpha = \frac{\tilde{\mu}_V^2}{\tilde{\sigma}_V^2} \tag{4.29}$$

und

$$\lambda = \frac{\tilde{\mu}_V}{\tilde{\sigma}_V^2} . \tag{4.30}$$

Soweit der Algorithmus. Im Anhang A.3 wird er zur Veranschaulichung auf ein einfaches Beispiel angewendet.

### 4.4.3 Ergebnisse

Der in Abschnitt 4.4.2 beschriebene Algorithmus wurde im Rahmen der Fortentwicklung des Tools TESter implementiert [Moh99] und evaluiert. Als typisches Beispiel für die damit erzielten Resultate zeigt Abbildung 4-24 die Vordergrundsequenz eines automatisierten Modells der Referenzkurve. Wir sehen, daß die wesentlichen Eigenschaften wie Periodizität und gute Übereinstimmung der Autokorrelationsfunktionen auch hier erfüllt sind, wenngleich sich das Aussehen der Simulationskurve von Periode zu Periode leider nicht allzusehr verändert. Als

zentraler Vorteil des gewählten Automatisierungsansatz bleibt jedoch in jedem Fall festzuhalten, daß sich der Aufwand zur Generierung eines überzeugend parametrisierten Modells (welches die eingangs aufgestellten Kriterien (i) - (iii) also hinreichend gut erfüllt) damit signifikant verkürzen läßt.

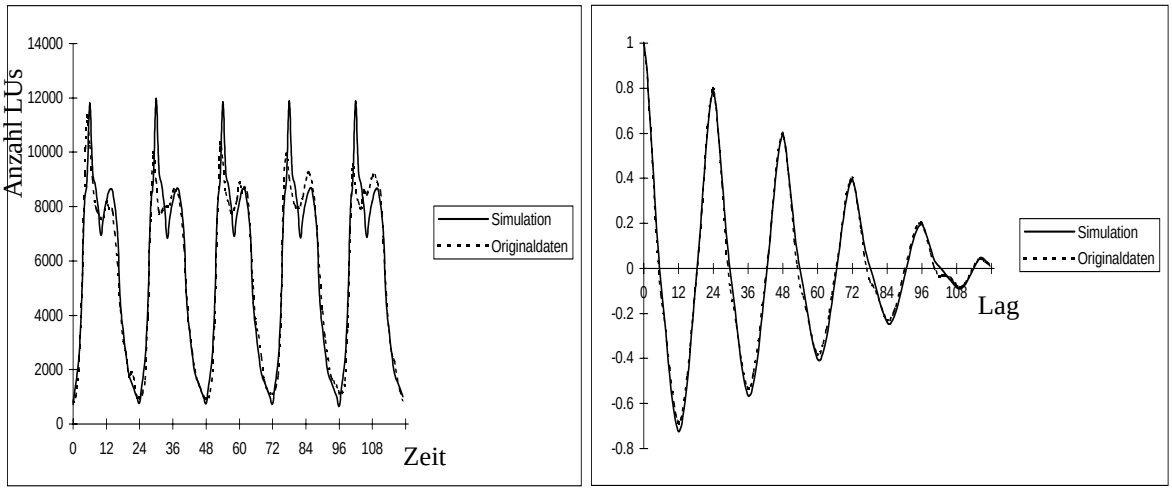


**Abbildung 4-24:** Modellierung der Referenzkurve mit dem automatisierten TES-Verfahren und zugehörige Autokorrelationsfunktionen

## 4.5 Fazit

Die in diesem Kapitel untersuchte und fortentwickelte TES-Methode wurde ursprünglich zur Modellierung von Zeitreihen im Hinblick auf Randverteilung, Autokorrelationsstruktur und Kurvenverlauf entwickelt. Nach ausführlicher Darstellung der mathematischen Grundlagen sowie der Parametrisierung des Standard-Verfahrens wurde das Konzept der Verallgemeinerten Stitching-Funktion eingeführt und die Konsequenzen für die Verzerrungs-Transformation gezogen. Am Beispiel der Referenzkurve wurde die dadurch erreichbare Verbesserung der Modellierung deutlich gemacht. Die Anwendung des neuen Verfahrens auf MPEG-codierten Videoverkehr erlaubte es, im Gegensatz zu bisherigen Ansätzen die unterschiedlichen Frame-Typen in ein gemeinsames Modell zu integrieren. Schließlich wurde noch ein Algorithmus zur automatisierten Parametrisierung der TES-Modelle für periodischen Verkehr angegeben und validiert, der eine deutliche Vereinfachung gegenüber bisherigen Automatisierungsmethoden bedeutet. Damit hat sich das verallgemeinerte TES-Modell als Verfahren erwiesen, das den in Kapitel 2 ursprünglich formulierten Anforderungen für einen Einsatz in einer GSM-Systemtest-Umgebung genügt. An der Realisierung wird im Rahmen eines entsprechenden Projekts (vgl. [Hak99]) der Ericsson Eurolab Deutschland GmbH derzeit gearbeitet.

KAPITEL 5

# *Ergänzende Bemerkungen zur TES-Modellierung*

## 5.1 Offene Fragen

Im vorhergehenden Kapitel wurde ausführlich dargestellt, welche neuen Möglichkeiten sich durch die Einführung der Verallgemeinerten Stitching-Funktion für die Verkehrsmodellierung mit TES ergeben. Aus den verbleibenden offenen Fragen werden im folgenden Abschnitt drei näher angerissen, um so einige Richtungen für weitergehende Untersuchungen anzudeuten.

### 5.1.1 Iterative Automatisierung von TES

In Abschnitt 4.4 wurde ein einfaches Verfahren vorgestellt, das eine automatisierte Anpassung der Parameter eines TES-Modells an die zu simulierende Situation (z.B. die Referenzkurve) ermöglicht. Das Vorgehen läuft dabei grundsätzlich in folgenden drei Schritten ab:

- Zunächst wird aus den vorliegenden empirischen Daten eine geeignete Verallgemeinerte Stitching-Funktion (GSF) bestimmt. Dies kann etwa durch Ermittlung eines durchschnittlichen Kurvenverlaufs der Referenzkurve o. ä. geschehen.
- Sodann wird die Referenzkurve als Realisierung eines TES-Modells mit der entsprechenden GSF aufgefaßt und durch “Zurückrechnen” die dieser Annahme zugrundeliegende Innovationssequenz und damit letztlich die entsprechende Innovationsdichte ermittelt.
- Der abschließende Schritt besteht dann in der Approximation dieser Innovationsdichte mit Hilfe einer geeigneten Gamma-Verteilung.

Dieses Vorgehen bietet zwei Ansatzpunkte für eine denkbare Verbesserung der Modellqualität. Zum einen wurde bislang die Wahl der GSF anhand naheliegender heuristischer Kriterien vorgenommen, zum anderen erscheint auch die Wahl der Klasse der Gammaverteilungen als geeignete Kandidaten für eine Approximation der zurückgerechneten Innovationsdichte in

gewissem Sinne als willkürlich. Die bekannten Eigenschaften der Gamma-Verteilungen können diese Wahl zwar rechtfertigen [Moh99], aber der ursprüngliche Vorschlag, eine stückweise gleichverteilte Innovationssequenz zu verwenden [JM92a], kann trotz erhöhter Parameterzahl zu vergleichbaren Ergebnissen führen [E-H97].

Vor allem aber sei an dieser Stelle auf ein mögliches Ausnutzen des Zusammenspiels zwischen diesen beiden Faktoren hingewiesen. Denn die Auswahl der GSF hat offensichtlich einen fundamentalen Einfluß auf die aus dem Zurückrechnen resultierende Innovationsdichte, insbesondere läßt sich feststellen, daß die Innovationsdichte umso einfacher angesetzt werden kann, je komplexer die Verallgemeinerte Stitching-Funktion gewählt wird. Daher erscheint die Konstruktion eines Iterationszyklus denkbar, der im mehrfachen Durchlaufen der eingangs dargestellten drei Schritte besteht, wobei jeweils aus dem Ergebnis der Approximation der Innovationsdichte eine Veränderung der GSF abgeleitet wird, die ihrerseits zu einer neuen (und hoffentlich einfacheren) Innovationsdichte führt usw., bis schließlich eine gewisse Optimierung des Modells erreicht wird, und zwar in dem Sinn, daß eine hinreichend einfache (und daher gut zu approximierende) Innovationsdichte im Zusammenspiel mit der parallel dazu entwickelten GSF die befriedigende Modellierung der vorliegenden Referenzkurve erlaubt.

### 5.1.2 Formale Behandlung des Kriteriums der visuellen Ähnlichkeit

In diesem Zusammenhang stoßen wir auf eine weitere bislang ungelöste Frage: Die “befriedigende” Modellierung einer Referenzkurve hängt im wesentlichen davon ab, in welchem Maße die in Abschnitt 4.1.1 eingeführten drei Kriterien von der modellierten Kurve erfüllt werden. Läßt sich nun die Erfüllung der Kriterien irgendwie quantitativ messen? Die Antwort fällt im Fall der ersten beiden Kriterien nicht so schwer: Die Übereinstimmung der beiden Randverteilungen wie auch die Äquivalenz der beiden zu vergleichenden Autokorrelationsfunktionen ist numerischer Behandlung zugängig. Für die Autokorrelationsfunktionen kann man beispielsweise (analog zu (4.24)) das oft verwendete Gaußsche Kriterium der Summe der (evtl. noch gewichteten) quadratischen Abweichungen der einzelnen Lags als Maß für die Übereinstimmung heranziehen; im Falle der Randverteilungen würde dem ein Integral über die quadrierte Differenz der Verteilungsdichten entsprechen. Insbesondere kann man feststellen, daß Kriterium (i) bzw. (ii) strikt erfüllt ist, wenn Summe bzw. Integral jeweils verschwinden (Randverteilungen bzw. Autokorrelationsfunktionen also exakt übereinstimmen).

Schwieriger stellt sich die Frage im Falle von Kriterium (iii) dar. Zu welchem Grade die Referenzkurve und die modellierte Kurve in ihrem Verlauf übereinstimmen, wurde bislang mehr oder weniger dem Gefühl bzw. der subjektiven Einschätzung des Modellierers überlassen (selbst der Automatisierungsvorschlag von [JM95] resultiert in einer größeren Anzahl von Vorschlägen für das gesuchte Modell, aus denen der Modellierer schließlich manuell die endgültige Auswahl trifft). Gerade im Falle von periodischen Kurven wie der in unserem Fall betrach-

teten Abbildung 2-3 ist aber auch hier eine formale Aussage über die Erfüllung oder Nichterfüllung des Kriteriums denkbar.

Die im folgenden hierzu vorgeschlagene Idee beruht auf der Annahme, daß der Verlauf der Referenzkurve durch zufällige (und daher von Periode zu Periode unterschiedliche) Abweichungen von einer strikt periodisch schwingenden “Basiskurve” zustandekommt. Diese Basiskurve läßt sich etwa durch Mittelung über die einzelnen Perioden bilden (wie wir das zur Ermittlung von Verallgemeinerten Stitching-Funktionen durchgeführt haben, vgl. Abschnitt 4.3.2). In diesem Fall läßt sich dann jeder einzelne Meßpunkt der Referenzkurve interpretieren als Realisierung einer Zufallsvariable, deren Erwartungswert der entsprechende Wert der Basiskurve ist. Seien z.B. $X_4, X_{28}, X_{52}, \ldots, X_{4+24N}$ die Werte der in Abbildung 2-3 zu beobachtenden Minima (Montag, Dienstag etc. jeweils um 4 Uhr morgens). Dann lassen sich durch Mittelung über diese Werte die entsprechenden Basiskurvenwerte

$$\xi_4 = \xi_{28} = \ldots = \xi_{4+24N} = \frac{1}{N+1} \sum_{i=0}^{N} X_{4+24i} \tag{5.1}$$

ermitteln. Außerdem läßt sich auf die übliche Weise die Varianz der Zufallsvariablen:

$$\sigma_4^2 = \sigma_{28}^2 = \ldots = \frac{1}{N} \sum_{i=0}^{N} (X_{4+24i} - \xi_4)^2 \tag{5.2}$$

schätzen. Dies läßt sich analog für alle Werte $(\xi_i, \sigma_i^2)$, $i = 1, 2, \ldots, 24$ durchführen.

Seien nun $\tilde{X}_1, \tilde{X}_2, \tilde{X}_3, \ldots$ die entsprechenden Werte, die das TES-Modell der Referenzkurve liefert. In diesem Fall kann man von strikter Erfüllung des Kriteriums (iii) sprechen, wenn für alle $i = 1, 2, \ldots, 24$

- der Erwartungswert von $\tilde{X}_i, \tilde{X}_{i+24}, \tilde{X}_{i+48}, \ldots$ jeweils gleich $\xi_i$ und
- die Varianz von $\tilde{X}_i, \tilde{X}_{i+24}, \tilde{X}_{i+48}, \ldots$ jeweils gleich $\sigma_i^2$ ist.

Weicht der Erwartungswert systematisch von dem Basiskurvenwert ab, so schlägt sich dies auch in der Randverteilung von Kriterium (i) nieder. Ist die Varianz größer als das zugehörige $\sigma_i^2$, so gebärdet sich das TES-Modell unregelmäßiger als die Referenzkurve, während bei zu kleiner Varianz das Modell der Referenzkurve zu eng folgt; in diesem Fall stellt sich die modellierte Kurve als zu regelmäßig dar, weist also *zuwenig* zufällige Abweichungen von der (gemittelten) Basiskurve auf.

Durch Vergleich von $E(\tilde{X}_i)$ mit $\xi_i$ und $Var(\tilde{X}_i)$ mit $\sigma_i^2$ $\forall\, i = 1, 2, \ldots, 24$ läßt sich also ohne großen Aufwand feststellen, ob die zufälligen Schwankungen der empirischen Referenzreihe in ausreichendem, aber nicht übertriebenem Maße von der Modellkurve wiedergegeben

werden. Natürlich sind neben diesem Vorschlag auch andere Ansätze denkbar und bilden Gegenstand aktueller und zukünftiger Untersuchungen.

### 5.1.3 Auflösungsqualität des Modells

Zum Schluß sei kurz auf einen weiteren Aspekt hingewiesen, der ebenfalls noch genauerer Betrachtung bedarf. Die Referenzkurve von Abbildung 2-3 beruhte auf Messungen, die im Stundenabstand vorgenommen wurden, woraus sich automatisch eine Periodenlänge von 24 ergibt. Die Frage liegt nahe, wie denn beim Vorliegen feinerer Meßdaten zu verfahren ist. Grundsätzlich lassen sich größere Periodenlängen natürlich stets über eine entsprechend komplexere Verallgemeinerte Stitching-Funktion angeben, allerdings um den Preis immer deterministischerer Innovationsdichten. Ein Ausweg hieraus bestünde in einem hybriden Ansatz, wonach das TES-Modell zur Grobmodellierung dient (also etwa basierend auf den vorliegenden stündlichen Messungen) und auf die resultierende Kurve ein normalverteiltes weißes Rauschen aufmoduliert wird. Man könnte aber auch versuchen, innerhalb des TES-Paradigmas zu verbleiben und beispielsweise wie gehabt ein grobes TES-Modell (für die stündlichen Meßergebnisse) erstellen und geeignet interpolieren, um sodann die Differenzen der feineren Meßergebnisse bezüglich der groben (interpolierten) Kurve als Grundlage für ein neues, in der Regel nicht mehr periodisches, TES-Modell zu verwenden. Der Vorteil hierbei liegt wiederum in der möglichen Anpassung von empirischer und modellierter Autokorrelationsfunktion, denn es ist anzunehmen, daß auch die Differenzen der Feinmessung zum interpolierten Grobmodell nicht unabhängig voneinander, sondern in gewissem Maße untereinander korreliert sind. Die Wahl zwischen diesen beiden sowie weiteren denkbaren Ansätzen wird allerdings stets von den jeweils zugrundeliegenden Meßdaten abhängen, weshalb der Einzelfall unter Umständen noch detailliertere Untersuchungen erforderlich machen kann.

Einen ganz anderen Weg schlägt [Hak99] vor. Anstatt die mit einer höheren Meßauflösung einhergehenden Ausschläge als statistische Schwankungen kumulierter Daten anzusehen, wird hier das Verhalten zugrundeliegender stochastischer Punktprozesse als ausschlaggebend angesehen. Die Modellierung des Kurzzeitverhaltens setzt daher bei der Modellierung der einzelnen Ankunftszeitpunkte an. Dies ist am einfachsten durch die Erzeugung eines inhomogenen Poissonprozesses mit unterschiedlichen Ankunftsraten in den verschiedenen Zeitabschnitten möglich. Liegt überdies burstartiges Verhalten der Ankünfte vor, so stellen sich Markov-modulierte Poissonprozesse (MMPP, vgl. Abschnitt 2.3.2) als geeignet für die Modellierung der Ankunftszeitpunkte heraus. Die Anzahl der Ankünfte innerhalb eines Zeitintervalls, die einen zentralen Parameter des entsprechenden Punktprozesses darstellt, wird dagegen stets mit Hilfe eines (gröberen) TES-Modells bestimmt. Neben der theoretischen Fundierung dieses Vorschlags finden sich in [Hak99] auch Algorithmen und Verfahren zur Parameterschätzung angegeben, ohne daß jedoch eine Implementation vorgenommen wurde, anhand derer die praktische Anwendbarkeit dieses Ansatzes validiert werden könnte. Auch an dieser Stelle tut daher eine weiterführende Untersuchung not.

## 5.2 Ausblick: Ein TES-Modell für selbstähnlichen Verkehr?

Zum Abschluß dieses Teils wenden wir uns nun noch einem weiteren Anwendungsgebiet der TES-Methode zu. Nachdem im vorigen Kapitel ausführlich gezeigt wurde, wie durch entsprechende Formgebung der Verallgemeinerten Stitching-Funktion eine zufriedenstellende Modellierung der Referenzkurve inklusive ihres äußerlichen Verlaufs erreicht werden kann, soll in zugegebenermaßen etwas spekulativer Art und Weise noch kurz die Frage aufgeworfen werden, ob neben den Maxima und Minima einer empirisch ermittelten Kurve vielleicht auch noch andere charakteristische Eigenheiten durch geeignete Anpassung der GSF berücksichtigt werden können. Hierbei richtet sich der Blick vor allem auf ein Gebiet der Verkehrsmodellierung, das seit der grundlegenden Arbeit von [LTWW94] wie kaum ein anderes für Aufsehen und Faszination gesorgt hat [WTE96]: die Modellierung von selbstähnlichen Verkehrsstrukturen.

Die TES-Methode zur Modellierung von selbstähnlichem Verkehr zu verwenden erscheint auf den ersten Blick schon deshalb nicht völlig abwegig, da sich die Selbstähnlichkeit eines stochastischen Prozesses grundsätzlich über die Form seiner Autokorrelationsfunktion definiert, auf deren Modellierung TES ja ausgerichtet war. Sei nämlich $(X_t)_{t = 1, 2, \ldots}$ ein stochastischer Prozeß, dessen Autokorrelationsfunktion für Lags $\tau \to \infty$ die Form

$$r(\tau) \propto \tau^{-\beta} \cdot L(t) \tag{5.3}$$

aufweist, wobei $\beta \in (0,1)$ ist und $L(t)$ hinreichend wenig variiert und der Einfachheit halber als konstant angenommen wird. Dann definiert man für jedes $m = 1, 2, \ldots$ einen neuen stochastischen Prozeß

$$(X_j^{(m)})_{j = 1, 2, \ldots} \text{ mit } X_j^{(m)} = \frac{X_{(j-1)m+1} + X_{(j-1)m+2} + \ldots + X_{jm}}{m}, \tag{5.4}$$

also den Mittelwert von $m$ aufeinanderfolgenden Realisierungen des ursprünglichen Prozesses (das Betrachten von Prozessen $X^{(m)}$ hat für kleiner werdendes $m$ also einen "Zoom-Effekt" zur Folge, der Prozeß erscheint in immer höherer Auflösung). $X^{(m)}$ besitzt seinerseits eine Autokorrelationsfunktion $r^{(m)}(\tau)$, und (*exakte*) *Selbstähnlichkeit* ist definiert als Invarianz dieser Autokorrelationsfunktion unter der Zoom-Transformation (5.4), wenn also gilt:

$$r^{(m)}(\tau) = r(\tau) \qquad \forall m, \tau = 1, 2, \ldots . \tag{5.5}$$

Der Exponent $\beta$ in (5.3) wird über $H = 1 - \beta/2$ in den sog. Hurst-Parameter transformiert, der in gewissem Sinn den "Grad" der Selbstähnlichkeit charakterisiert [Ro96].

Selbstähnlichkeit läßt sich auf einige verschiedene Arten nachweisen ([TTW95], [Fas98]). Wir beschäftigen uns nur mit der am weitesten verbreiteten, der sog. *Aggregated Variance Method*. Hierbei schätzt man für $m = 1, 2, \ldots$ die jeweilige empirische Varianz des Prozesses

$(X_j^{(m)})_{j=1,2,\ldots}$ und trägt ihren Logarithmus schließlich gegen den Logarithmus von $m$ auf. Falls Selbstähnlichkeit vorliegt, ergibt dies eine lineare Funktion mit Steigung $-\beta \in (0,-1)$, während eine Steigung von exakt $-1$ charakteristisch für einen Poissonprozeß ist.

Vor diesem Hintergrund drängt sich der Versuch auf, ob man möglicherweise durch Einsatz einer *fraktalen Stitchingfunktion* in TES eine selbstähnliche Vordergrundsequenz erzeugen könnte. Als Kandidat hierfür wurde die Klasse der *selbstaffinen Kurven* [Fal90] ausgemacht. Sei hierzu $S_i$ $(1 \le i \le k)$ eine affine Transformation, die in Matrix-Notation bezüglich der Koordinaten $x$ und $t$ durch

$$S_i\begin{pmatrix} t \\ x \end{pmatrix} = \begin{bmatrix} \frac{1}{k} & 0 \\ a_i & c_i \end{bmatrix} \cdot \begin{pmatrix} t \\ x \end{pmatrix} + \begin{pmatrix} \frac{(i-1)}{k} \\ b_i \end{pmatrix} \tag{5.6}$$

repräsentiert ist. Mit anderen Worten: man unterteile das Einheitsintervall in $k$ gleiche Teile. Dann kontrahiert $S_i$ das Einheitsintervall um den Faktor $1/k$ und bildet es auf das $i$-te neue Intervall ab, während $c_i$ die Kontraktion in x-Richtung beschreibt (o.B.d.A. $1/k \le c_i \le 1$). Eine fraktale selbstaffine Kurve entsteht durch wiederholtes Ersetzen der einzelnen Linienelemente durch ihre affine Transformation, wie in Abbildung 5-1 für $k = 3$ durch die ersten drei Iterationen $(n = 1, 2, 3)$ und das Resultat für eine sehr hohe Iterationszahl $n \to \infty$ veranschaulicht wird (vgl. [Rei98a]).

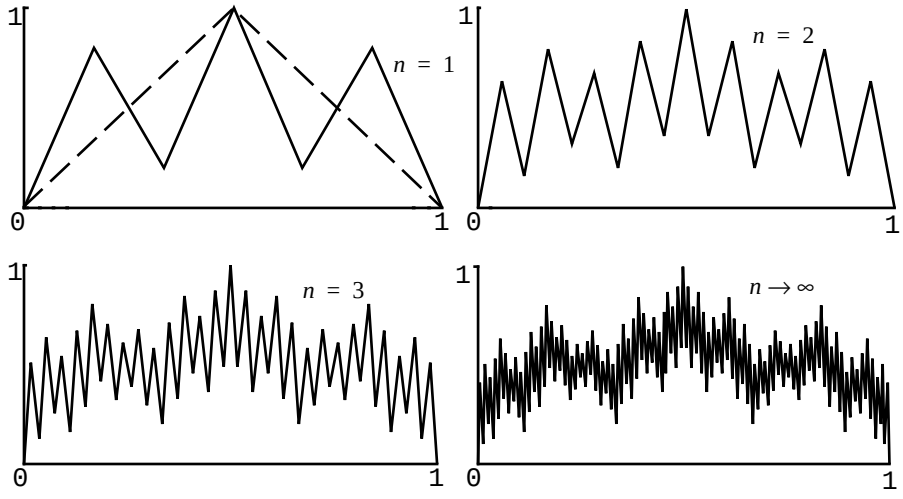


**Abbildung 5-1:** Auf dem Weg zu einer fraktalen selbstaffinen Kurve

Die Verwendung einer derartigen Kurve (mit Iterationsordnung 10) als Verallgemeinerte Stitchingfunktion liefert im Falle einer driftlosen Hintergrundsequenz ($\delta = 0.05$) tatsächlich eine Vordergrundsequenz mit selbstähnlichen Strukturen. Abbildung 5-2 zeigt neben der verwendeten ursprungssymmetrischen Innovationsdichte und der Hintergrundsequenz unten rechts die daraus resultierende Vordergrundsequenz und in der Mitte und links die daraus über (5.4) für

$m = 10$ bzw. $m = 100$ erhaltenen "entzoomten" Sequenzen $X_j^{(m)}$. Offensichtlich bleibt hierbei die "Burstiness" der Kurve trotz Aggregation weitgehend erhalten, was kennzeichnend für selbstähnliche Prozesse ist. Deshalb wurde auch noch die erwähnte Aggregate-Variance-Methode auf die Sequenzen $X_j^{(m)}$ mit $m = 2, 5, 10, 20, \ldots, 1000$ angewandt und ergab die Kurve von Abbildung 5-3. Wenn man berücksichtigt, daß die zusätzlich eingezeichnete Linie mit Steigung $-1$ das für nicht-selbstähnliche Sequenzen obligatorische Ergebnis ist, kann man festhalten, daß zumindest über die betrachteten Größenordnungen hin ein tatsächlich ein gewisses Maß an Selbstähnlichkeit vorliegt.

Dieses Ergebnis ist allerdings sehr wohl cum grano salis zu genießen, da seine Reproduktion typischerweise mit großem Aufwand zur Feinjustierung der zugrundeliegenden Hintergrundsequenz verbunden ist. Ein weiteres grundsätzliches Problem dabei ist die Bestimmung der Randverteilung der fraktalen GSF für hohe Iterationsordnungen (insbesondere im Limesübergang).

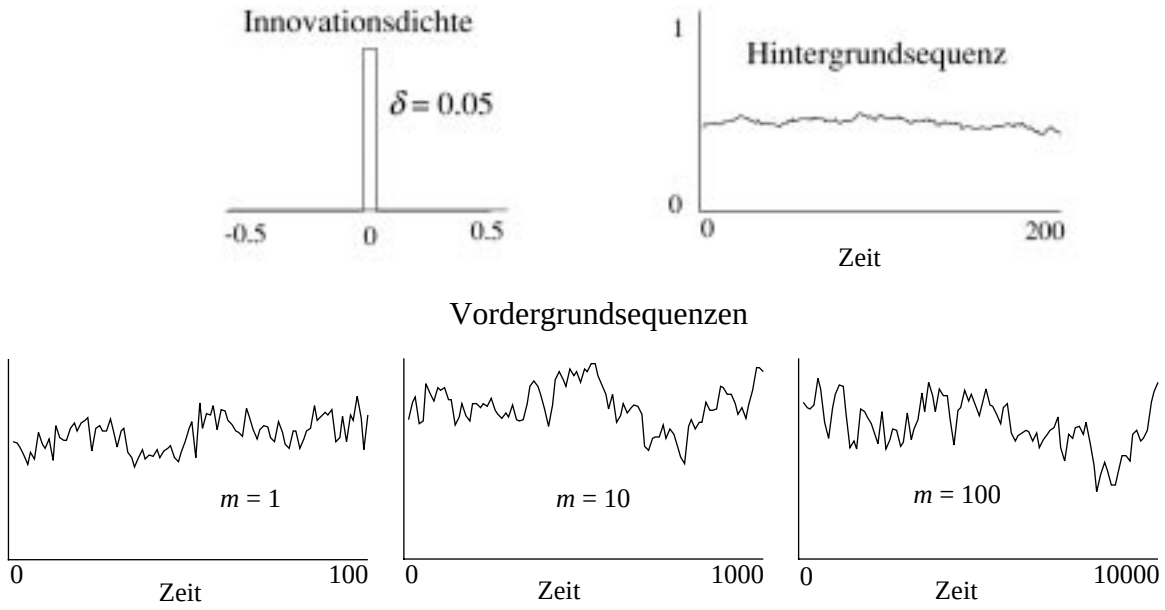


**Abbildung 5-2:** Fraktales TES: Innovationsdichte, Hintergrundsequenz und gezoomte Vordergrundsequenzen (jeweils um Faktor 10)


**Abbildung 5-3:** Aggregate Variance-Plot für die Kurven aus Abbildung 5-2

Schließlich ist noch zu beachten, daß typischerweise von Selbstähnlichkeit erst die Rede ist, wenn die beschriebenen Effekte über fünf und mehr Größenordnungen unverändert feststellbar sind, während unsere Experimente in der Regel nach mehr als drei Größenordnungen eine allmähliche Annäherung der Steigung an die Winkelhalbierende zeigten (dies deutet sich auch am rechten Ende der Kurve in Abbildung 5-2 schon an). Andererseits läßt sich diesem Einwand auch anders begegnen, wenn man bedenkt, daß Selbstähnlichkeit im Sinne ihrer mathematischen Definition (5.4) keine Begrenzung von $m$, also im wesentlichen der erwähnten Größenordnung vorsieht, während alle bekannten praktischen Messungen früher oder später ebenfalls eine Annäherung der Steigung im Aggregate Variance-Plot an die Winkelhalbierende zeigen und damit im strengen Sinne eben doch nur *short-range dependent* statt *long-range dependent*, also selbstähnlich, sind. Daher wurde in [Rei98a] vorgeschlagen, im Sinne einer besseren Erfassung der praktischen Realitäten den Begriff des *medium-range dependent traffic* einzuführen und damit Sequenzen zu bezeichnen, bei denen Selbstähnlichkeit nur über eine begrenzte Anzahl von Größenordnungen zu beobachten ist. Der Ausbau dieses Konzepts ist Gegenstand aktueller Arbeiten (u.a. [Bro98], vgl. auch [RLeB97]).

Mit diesen Überlegungen sind wir schließlich am Ende dieses ersten Teils an einer anderen Stelle der anfangs beschriebenen Kluft zwischen Theorie und Praxis angekommen. Wieder einmal stellen wir fest, daß sich die Überprüfung strenger Mathematik an den tatsächlich vorliegenden Gegebenheiten und in der Folge die Anpassung der Modellbildung und ihrer Methodik an die vielfältigen praktischen Erfordernisse als ständig neue Herausforderung präsentiert. Dies wird auch am nun folgenden zweiten großen Themenkomplex dieser Arbeit deutlich werden, der sich mit der Suche nach dynamischen Tarifmodellen für das Internet auseinandersetzt.

KAPITEL 6

# *Preismodelle für Kommunikationssysteme: Übersicht und Herausforderungen*

## 6.1 Überblick und Klassifizierung

Nicht unerwartet hat der Internet-Boom der vergangenen Jahre mit seinem mehr oder weniger exponentiellen Wachstum in verschiedensten Bereichen zu einer wahren “Goldgräberstimmung” geführt. Seit insbesondere die Finanzierung des Internet selbst von staatlichen Stellen mehr und mehr in die Hände kommerziell geführter Unternehmungen übergeht, hat die Frage nach internetspezifischen Tarifmodellen rapide an Interesse gewonnen. Allerdings stellte sich heraus, daß der Weg zu einer Lösung, die sich allgemein durchsetzen könnte, noch weit ist. Immerhin existieren bereits eine beachtliche Anzahl verschiedener Vorschläge, wie sie nachfolgend in einer Art Bestandsaufnahme zusammengefaßt werden. Im Anschluß daran erfolgt in den Kapiteln 7 und 8 eine vertiefte Untersuchung und Weiterentwicklung zweier vielversprechender Ansätze, die auf dem Weg weg vom starren Schema der heutigen “Flat Fee”-Tarife mit ihren monatlich fixen Gebühren hin zur Dynamik volumen- bzw. auktionsbasierter Preismechanismen liegen.

### 6.1.1 Ein Blick in die Praxis

Es sind nur wenige Preismodelle, die im bisherigen Internet eine weitere Verbreitung und Anwendung gefunden haben, insbesondere aus dem Blickwinkel von Dienstanbietern (Internet Service Providers, ISPs) betrachtet. Der Internet-Zugang kostet den Nutzer meist eine fixe monatliche Gebühr, die eine Nutzung in unbeschränkter Weise (“Flat Fee”-Tarif) oder doch zumindest für eine gewisse begrenzte Zeit beinhaltet. In letzterem Fall schlägt sich zusätzliche Nutzung in einer meist stundenweise abgerechneten Gebühr nieder. Volumenabhängige Gebühren wurden in der Vergangenheit von einigen Providern verwendet, scheinen aber in jüngster Zeit an Popularität eingebüßt zu haben. Derzeit beginnt sich sogar ein Trend in die entgegengesetzte Richtung zu etablieren, bei dem das Internet-Surfen völlig umsonst ist, wenn sich der Kunde im Gegenzug vertraglich an einen bestimmten Telefonnetz-Anbieter bindet.

Die monatliche Gebühr für einen Internetanschluß hängt normalerweise von der Bandbreite des Zugangs ab. In vielen Fällen ist es dem Kunden überlassen, für die physikalische Verbindung hin zum nächstgelegenen POP (Point-of-Presence) des ISP (Internet Service Provider) zu sorgen. Meistens wird es sich dabei um eine Mietleitung handeln, was eine weitere bandbreiten- und entfernungsabhängige Gebührenerhebung zur Folge hat (die oftmals den Hauptanteil der Kosten ausmacht). Bei festen Internetverbindungen sind volumenabhängige Gebühren noch weiter verbreitet, insbesondere außerhalb von Nordamerika. Sie bestehen gewöhnlich aus einer fixen Komponente, die von der Zugangskapazität abhängt, plus zusätzlichen Gebühren pro übertragenem Megabyte (oft mit hinreichendem Mengenrabatt). Die Gewichtung dieser beiden Anteile, Preis-pro-Volumen-Kurven und weitere Parameter (wie z. B. die Frage, ob sich beide Verkehrsrichtungen in der Volumengebühr niederschlagen) können von Anbieter zu Anbieter erheblich variieren. Meistens aber beinhaltet die fixe Komponente zumindest einiges an Datenvolumen, das umsonst übertragen werden kann.

In letzter Zeit ist als weitere Variante vereinzelt eine “Burst”-Gebühr ins Gespräch gekommen, wobei der ISP periodisch (z. B. stündlich) das übertragene Datenvolumen mißt. Monatlich werden dann alle Messungen nach der Größe geordnet, dann wird ein gewisser Teil (etwa die obersten 5%) verworfen, um ungewöhnlich hohe Peaks auszusondern, und die verbleibende höchste Messung dient als Grundlage für die Übertragungsbandbreite, auf der die erhobene Gebühr basiert. Darüberhinausgehende Ansätze von nutzungsbasierten Tarifen, etwa differenziert nach Entfernung oder Uhrzeit, sind allerdings kaum anzutreffen, mit einer berühmt gewordenen Ausnahme, nämlich den volumenbasierten Gebühren für Bottleneck-Verbindungen wie etwa die transozeanischen Verbindungen der Forschungsnetze von Neuseeland [Bro97] und Großbritannien [Rog98].

### 6.1.2 Zur Klassifizierung von Preismodellen für das Internet

Eine Klassifizierung von Preismodellen (wie in [RLS99] vorgeschlagen) sollte – neben der Frage nach der Praxisrelevanz eines Vorschlags – mindestens vier weitere grundlegende Aspekte spezifizieren, nämlich (a) die Abhängigkeit der tarifierten Dienste von der technischen Charakteristik des zugrundeliegenden Netzes, (b) die denkbaren Komponenten, aus denen sich ein Tarif zusammensetzen kann, (c) die meßbaren Verkehrsparameter, die für die Tarifbildung zur Verfügung stehen, und (d) die von einem bestimmten Preismodell verfolgte Absicht.

Vor diesem Hintergrund lassen sich folgende drei Hauptdimensionen identifizieren, die in einem gewissen Sinne als orthogonal zueinander aufgefaßt werden können und damit gewissermaßen den Raum aufspannen, um die Diskussion von Preismodellierung und Tarifgestaltung zu ermöglichen [SRL99]: (1) die *technische Dimension*, (2) die *ökonomische* Dimension, und (3) die *Forschungsdimension*. Es stellt sich heraus, daß jede dieser Dimensionen obendrein aus zwei Schichten besteht: einer höheren Schicht im Sinne der Repräsentation einer abstrakt gehaltenen Sicht, und einer tieferen Schicht, die sich mit konkreten Aspekten beschäftigt.

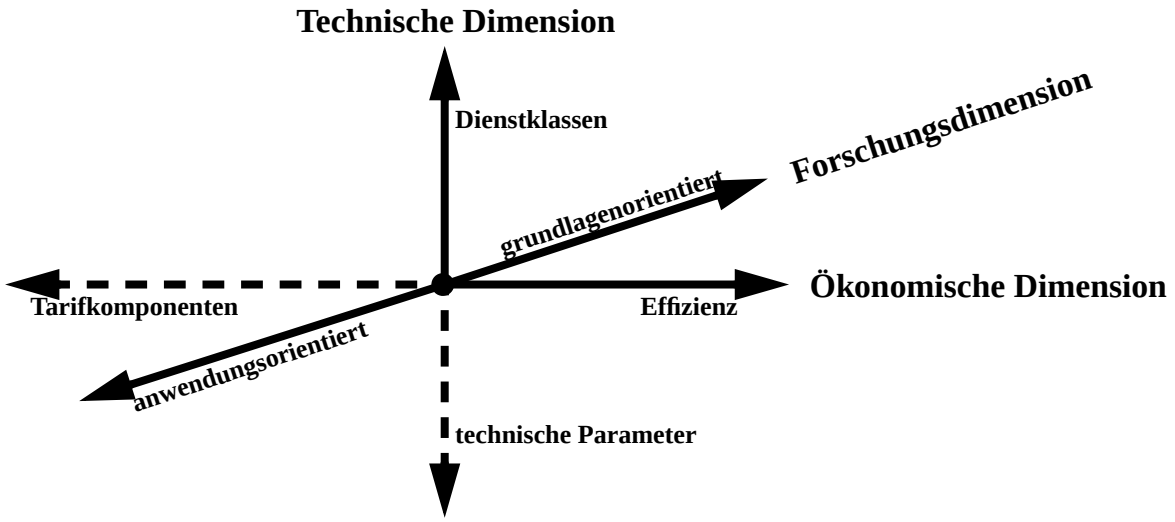


**Abbildung 6-1:** Dimensionen zur Klassifizierung von Preismodellen

Im folgenden werden wir kurz auf die einzelnen Subdimensionen und ihre Bedeutung eingehen, bevor wir uns eingehender mit den bereits vorliegenden Ansätzen beschäftigen.

### Dienstklassen

Zunächst findet sich unter den relevanten Vorschlägen zur Preismodellierung eine offenkundige Unterscheidung zwischen verbindungsorientierten und verbindungslosen Ansätzen. In der Geschichte der Internet-Tarifierung wurden *paketbasierte Modelle* als erste dazu genutzt, um eine bevorzugte Behandlung von (durch “Priority Bits” ausgezeichneten) Paketen zu ermöglichen, und zwar in Form von sogenanntem “relativen Best-Effort-Verkehr”.

Der *Integrated Services*-Ansatz erlaubt eine höhere Form von QoS-Garantien, indem er zu verbindungsorientierten Mechanismen übergeht. Durch die Verwendung von Reservierungsprotokollen (wie z.B. RSVP) wird es möglich, in Überlastsituationen neu hinzukommende Kunden abzublocken. Da hierbei jedoch jede Verbindung als individueller Flow behandelt und über eine eigene sog. “flowspec” charakterisiert wird, sind Skalierungsprobleme unausweichlich.

Daher ist seit kurzer Zeit ein weiterer Vorschlag in der Diskussion, der zumindest im Internet-Backbone diese Problematik zu lösen versucht. Statt die einzelnen Flows individuell zu spezifizieren, geht dieser sog. *Differentiated Services*-Ansatz davon aus, daß der ISP gewisse Dienstklassen anbietet, in welche die einzelnen Pakete anhand eines charakteristischen Parameters im IP-Header, dem PHB (Per-Hop-Behaviour), eingeordnet werden. Innerhalb einer Dienstklasse werden die Pakete dann unterschiedslos behandelt. Ansonsten ist *DiffServ* weiterhin auf IP-Technologie aufgebaut, also insbesondere verbindungslos, was die Eignung für Echtzeitanwendungen zumindest in Frage stellt.

### Technische Parameter

In dieser Dimension werden die Parameter erfaßt, die für die Verwendung in Tarif- und Gebührenmechanismen zur Verfügung stehen (oder zumindest zur Verfügung gestellt werden sollten). Die entsprechende Liste beginnt bei Prioritäts-Flags und Paketmarkierung und führt über Peak-, Nominal Bit- oder mittlere Flow-Raten und effektive Bandbreiten hin zu Parametern wie erwartete Pfad- oder Staukosten, dynamische Gebote pro Paket oder Einheitsressource (im Fall von Auktionen), um nur einige aufzuzählen.

### Tarifkomponenten

Die aus der Telekommunikation vertrauten Tarife bestehen im wesentlichen aus drei Grundelementen, nämlich Gebühren für Netzzugang, Verbindungsaufbau und Nutzung (in Form von Gesprächsdauer). Unterschiedliche Kombinationen dieser drei Elemente ergeben eine Fülle von Tarifmöglichkeiten wie Flat Fee, nutzungs-, reservierungs-, volumen-, dienstklassen- oder bandbreitenbasierten Tarifen u.v.m. Auf ähnliche Weise kann man bei der Internet-Tarifierung eine Anzahl grundlegender Preismechanismen unterscheiden (z. B. Flat Rate, nutzungssensitive oder volumenbasierte Bepreisung, Paket- oder Flow-Auktionen, Tarife basierend auf Dienst- oder Nutzerprofilen, Edge Pricing contra multilaterale Kontrakte, usw.).

### Effizienz

Schließlich ist noch die Frage zu berücksichtigen, was durch die Tarifgestaltung denn überhaupt erreicht werden soll. Tarife kann man nämlich für unterschiedliche Ziele verwenden: einmal im Hinblick auf Netzeffizienz, also Maximierung der Ressourcennutzung (wie Bandbreite, Speicherplatz etc.), andererseits aber auch zur ökonomischen Effizienz, d.h. im Hinblick auf den Nutzen für den Kunden. Folglich maximiert ein geeigneter Tarif entweder die Einnahmen des Anbieters (durch effiziente Aufteilung der Ressourcen sowie Zugangskontrolle) oder die Zufriedenheit des Nutzers. Hier kommen dann auch Fairneß-Aspekte oder die sog. “Incentive Compatibility” ins Spiel (mehr dazu in Kapitel 8).

### Forschungsrichtung

Ergänzend hierzu sind noch einige weitere Gesichtspunkte zu nennen. So hat es beispielsweise in der Regel einen wichtigen Einfluß auf die Gestaltung eines Tarifmechanismus, ob ein in der Forschung vorgeschlagenes Modell eher aus einer theoretisch oder aber einer praktisch orientierten Ecke kommt. Beide Richtungen können wichtige Grundlagen für die tatsächliche Realisierung eines entsprechenden Gebührensystems liefern, wobei jedoch festzuhalten bleibt, daß momentan der überwiegende Teil der relevanten Arbeiten auf sehr hohem theoretischen Niveau stattfindet.

#### Weitere Aspekte

Neben diesen allgemeinen Dimensionen sind für ein viables Preismodell noch eine Fülle weitere Gesichtspunkte einzubeziehen. Diese können von Applikationstypisierung (z.B. Burstiness) über technologische und wirtschaftliche Aspekte (Grenzkosten, Überlastgebühren, Responsive Pricing, schließlich auch die Frage, ob letztlich der Sender oder der Empfänger bezahlt) bis hin zu überaus praktischen Fragen (wie Transparenz, Vorhersagbarkeit, Praktikabilität, Annahme durch den Nutzer und Benutzerfreundlichkeit) reichen, ohne daß diese Liste Anspruch auf Vollständigkeit erhebt.

## 6.2 Ansätze und Probleme

Wenngleich die Frage nach einer Tarifierung von Internetdiensten ihre brennende Aktualität erst in allerjüngster Zeit im Gefolge des kometenhaften Aufschwung des Internets erhalten hat, so lassen sich bei einem Blick in die Wissenschaftsliteratur doch schon einige relativ weit gereifte Ansätze ausmachen. Die ersten Ideen am Anfang der 90er Jahre bevorzugten gewöhnlich eine Art einfaches Dienstprioriätsmodell, bevor im Jahr 1993 [MV93] die Idee einer Verwendung von Auktionsmechanismen einführte. 1995 formulierte [She95] dann ein wichtiges Modell, das erstmals auf dem IntServ-Ansatz basierte. Beide Arbeiten hatten auf den Fortgang der Forschung in den nächsten Jahren eine enorm stimulierende Wirkung (vgl. insbesondere die Papersammlung in [MB97]), aber es dauerte noch bis etwa 1997/98, bevor man dazu überging, erste Schritte hin zum Design von auch im wirklichen Leben verwendbaren Gebührenmechanismen zu unternehmen, meistens auf der Grundlage des RSVP-Protokolls.

In diesem Abschnitt werden in der gebotenen Kürze einige der hierfür relevantesten Internet-Preismodelle erläutert, wie sie im Lauf der vergangenen fünf Jahre Gegenstand der wissenschaftlichen Diskussion waren [RLS99]. Ausgehend von dem wichtigen Konzept des Edge Pricing liegt der Schwerpunkt vor allem auf diversen Auktionsansätzen, hierauf werden Nutzer- bzw. Service-Profile zur Sprache kommen, bevor wir uns eher praktisch orientierten Ansätze zuwenden werden.

### 6.2.1 Edge Pricing

Das grundlegende Konzept des Edge Pricings ([CESZ93], [SCEH96]) besteht darin, die Feststellung und Abrechnung der vom Nutzer für eine bestimmte Verbindung zu entrichtenden Gebühren an einer einzigen Stelle zu konzentrieren, nämlich beim ersten Netzanbieter entlang des entsprechenden Pfades, welcher durchaus auch Dienste weiterer ISPs nutzen kann (und in aller Regel auch wird). Die vom Kunden an den “Zugangs-ISP” bezahlte Gebühr muß also auch die Ausgaben für alle anderen mit dem Transport der entsprechenden Daten befaßten ISPs einschließen. Hauptvorteil dieses Konzeptes ist die Reduktion eines ansonsten für den

Ende-zu-Ende-Transport benötigten multilateralen Kontraktes zwischen allen beteiligten ISPs und dem Kunden auf eine Sequenz von bilateralen, was die Komplexität des Mechanismus' signifikant senkt und darüberhinaus eine deutlich höhere Transparenz für den Kunden zur Folge hat. Grundsätzlich spezifiziert hierbei der Kunde den maximalen Preis, den er als Sender bzw. Empfänger für eine Ende-zu-Ende-Verbindung zu zahlen bereit ist, außerdem noch eine Obergrenze für die Anzahl Hops. Diese Information kann als Teil eines Signalisierungsprotokolls, z. B. im RSVP-Header, übertragen werden. Üblicherweise findet bei der Preisbestimmung dann noch eine zweifache Approximation statt, wobei die derzeitigen Auslastungsbedingungen durch geeignete Schätzungen davon ersetzt werden (was z.B. zu tageszeitabhängigen Tarifen führt), ferner ersetzt man den Preis für den tatsächlich eingeschlagenen Pfad durch den Preis für den erwarteten Pfad.

### 6.2.2 Auktionsmechanismen

Die bahnbrechende Arbeit [MV93] geht aus von der Frage, ob und wie eine *ökonomisch effizient* gewählte Tarifstruktur dazu dienen kann, Überlastsituationen in den Griff zu bekommen, den Ausbau des Netzes an den richtigen Stellen zu fördern und die Ressourcennutzung zu optimieren (insgesamt also *technisch effizient* zu arbeiten). Da die Grenzkosten für den Pakettransport im wesentlichen verschwinden, solange die Ressource nicht ausgelastet ist, geben nutzungssensitive Preisschemata gute Kandidaten für einen Mechanismus zur Überlastkontrolle ab, weil sie die Frage einer effizienten Allokation von raren Ressourcen in einem wirtschaftlichen Kontext angehen. Hierbei ist zu betonen, daß der Schwerpunkt nicht auf dem Erzielen möglichst hoher Profite für den Netzbetreiber liegt, sondern auf der optimalen Nutzung der vorhandenen Ressourcen. Hier liegt der entscheidende Nachteil vieler heute populärer Tarifmodelle (wie in Abschnitt 6.1.1 kurz gestreift), die oft keine Anreize dafür vorsehen, überhöhte Peaks in Zeiten einer Netzüberlastung abzuflachen, wohingegen ein "ideales" Preisschema die Kosten widerspiegeln sollte, die ein Nutzer generiert, damit dieser fundierte Entscheidungen über die effiziente Ressourcennutzung treffen kann. Diese Kosten reichen von Fixkosten für die Infrastruktur des Netzes über Kosten für den Verbindungsaufbau und das darauffolgende Verschicken von Paketen bis hin zu "sozialen Kosten", die dadurch entstehen, daß bei der Nutzung einer überlasteten Resource die Pakete anderer Nutzer dadurch zum Warten verdammt werden.

Ein derartiges Tarifschema, bei dem Pakete nur dann Gebühren verursachen, wenn das Netzwerk aus- oder überlastet ist, kann mit Hilfe des "Smart Market"-Konzepts implementiert werden. Hierbei variiert der Preis für das Senden eines Pakets auf einer sehr kleinen Zeitskala, z.B. im Minuten- oder gar Sekundenbereich, und kann dadurch die momentane Auslastungssituation der Resource exakt widerspiegeln. Jeder Paketheader enthält ein sogenanntes "bid field", in dem der Absender ein Gebot für den Transport dieses Pakets abgibt, und das Paket wird übertragen, wenn dieses Gebot über den aktuellen Grenzkosten für diese Übertragung liegt. Das wichtigste Charakteristikum dieses Schemas ist, daß der Nutzer nicht etwa den in seinem

Gebot ausgedrückten Preis bezahlen muß, sondern vielmehr den (in der Regel darunter liegenden) aktuellen Marktpreis, also die besagten Grenzkosten. Diese sog. “Second-Bid”-Auktion (auch “Verallgemeinerte Vickrey-Auktion” genannt, vgl. [Vick61]) hat den Vorteil, daß – wie sich theoretisch beweisen läßt [LS97] – die für den Nutzer optimale Gebotsstrategie sehr einfach ist: wenn er als Gebot das angibt, was ihm der Transport des Pakets bzw. die Nutzung der Resource in Tat und Wahrheit wert ist, dann fährt er regelmäßig besser als wenn er versucht, mittels irgendwelcher Spekulationen den Markt zu überlisten. Für den Netzanbieter wiederum erlaubt das, ohne Zusatzaufwand an entscheidende Informationen über die Nutzer zu gelangen.

Eine Verallgemeinerung dieses Ansatzes findet sich in [M-M97]. Der Schwerpunkt dieser Arbeit liegt auf der Möglichkeit, Garantien für multiple QoS (speziell bei inelastischem Verkehr) abzugeben, wenn für die entsprechende Ressource im voraus Reservationen durchgeführt werden. Maximiert wird dabei die Summe der “Nutzer-Utilities” (vgl. hierzu Abschnitt 8.2.2), und zwar abstrakt gesehen durch Lösung eines Routing-Problems mittels Standard-Multicommodity-Flow-Techniken. Hierbei erlaubt der Begriff der “effektiven Bandbreite” [Kel91a] die Charakterisierung eines breiten Spektrums von Quellentypen mit Hilfe eines einzigen Parameters, der für die Reservation verwendet wird. Es läßt sich zeigen, daß sich das Routing und die Utility-Optimierung elegant entkoppeln lassen; die Lösung des entstehenden linearen Optimierungsproblems und vor allem des zugehörigen Dualproblems erlaubt eine Interpretation der Lösung als sogenannte “Spot-Preise”, also als Preise, die für jeden einzelnen Knoten angeben, wie teuer es ist, an dieser Stelle Verkehr in das Netz einzuspeisen bzw. aus dem Netz abzuziehen. Damit lassen sich die Grenzkosten des Systems für Verkehr von Knoten A nach Knoten B mit Hilfe von nur zwei Zahlen ausdrücken, nämlich dem Spot-Preis für A und dem für B; dies bedeutet, daß keine Routing-Information mehr benötigt wird. Allerdings erfordert die Bestimmung der optimalen Spot-Preise stets die Lösung des vollen zentralen Optimierungsproblems. Eine Dezentralisierung läßt sich nur erreichen, wenn die Nutzer bereit sind, ihre tatsächlichen “Utilities” wahrheitsgemäß offenzulegen, so daß eine Pareto-effiziente Allokation bestimmt werden kann. Wir haben oben bereits erwähnt, daß der “Smart Market” genau diese Eigenschaft gewährleistet und deshalb auch hier zum Einsatz kommen kann.

Ein verwandter Ansatz zur Verwendung von Auktionen, um die Entscheidungsfindung in paketvermittelnden integrierten Multiservice-Netzen zu dezentralisieren, kommt von [LS97]. Die Verallgemeinerung der Vickrey-Auktionen führt dabei zu einem Mechanismus, der sich als stabil, einfach, effizient und fair herausstellt. Im Gegensatz zum ursprünglichen “Smart Market”-Ansatz, der eindimensionale Gebote (Preis pro Paket) vorsieht und deshalb eine zentrale Bestimmung des Markträumungspreises basierend auf der expliziten Angabe der Utility-Funktionen durch die Nutzer erforderlich macht, benötigt dieser Ansatz die Möglichkeit, Ressourcen pro Flow zu reservieren. Daraus ergeben sich jetzt zweidimensionale Gebote (benötigte Menge an Ressource und Preis pro Ressourceneinheit), die eine Bestimmung des Räumungspreises anhand allein der Gebote ermöglichen. Ferner muß die Ressource nicht mehr

künstlich in kleine atomare Teile aufgeteilt werden (was entsprechend einen beträchtlichen Verlust an Flexibilität und Skalierbarkeit zur Folge hat), sondern Reservationen können für beliebige Teile der gesamten verfügbaren Ressourcen gemacht werden. In der entsprechenden spieltheoretischen Ausarbeitung dieses Ansatzes werden die Nutzerpräferenzen wiederum mittels Utility-Funktionen ausgedrückt, die für jeden Nutzer die individuelle Wertschätzung eines Menge/Preis-Vektors ausdrücken. Die daraus abgeleitete "Progressive Second Price"-Regel (PSP) verallgemeinert die Idee der Vickrey-Auktionen in ähnlicher Weise wie schon [M-M97], auch wenn dies wesentlich komplizierter ausgedrückt ist: der von mir als erfolgreichem Bieter zu bezahlende Preis pro Einheit bestimmt sich aus den Geboten aller anderen Mitspieler, wobei jedes dieser Gebote damit gewichtet wird, wie sehr die Reservation dieses Mitspielers durch das bloße Vorhandensein meines Gebotes beeinträchtigt wird. Somit bezahlt für jede infinitesimale Ressourcenmenge der Spieler, der sie erhält, den maximalen Preis, den ein Spieler, der sie gerade nicht erhält, zu zahlen bereit gewesen wäre. Diese Formulierung der Auktionsregel erlaubt die mathematische Ableitung einer Anzahl wünschenswerter Eigenschaften, insbesondere der Existenz eines fairen und effizienten Nash-Gleichgewichtes.

Zusammenfassend läßt sich festhalten, daß sich Second-Price-Auktionen als ein sinnvolles Konzept zur Bestimmung des aktuellen Marktpreises für ein überlastetes Netzwerk herausgestellt haben. Auf Fragen der tatsächlichen Realisierbarkeit solcher Vorschläge, insbesondere auch auf den dafür notwendigen enorm hohen technischen Aufwand, wird jedoch in den Arbeiten zu diesem Themenkreis meist nicht weiter eingegangen (vgl. [Vög98]). Deshalb werden wir in Kapitel 8 Vorteile wie Problematik solcher auktionsbasierter Ansätze im Hinblick auf ihre reale Einsetzbarkeit vertiefend diskutieren und uns hierbei insbesondere auf den Fall von Multiprovider-Szenarien konzentrieren, bei dem Auktionen für mehrere aneinanderzufügende Ressourcen, die aber verschiedenen Netzanbietern gehören, durchzuführen sind.

### 6.2.3 Profile und Klassen

[CF98] untersucht die Frage, ob und wie sich unterschiedliche QoS mit hoher Vorhersagbarkeit realisieren läßt, ohne das Best-effort-Konzept des IP-Protokolls aufzugeben. In Abkehr von der expliziten Reservation von Kapazität für den Nutzer wird im vorgestellten "Expected Capacity Framework" für jeden Nutzer ein Dienstprofil spezifiziert, woraufhin sich die einzelnen Anfragen jeweils in solche innerhalb des vorgegebenen Profils und solche außerhalb einteilen lassen. Beide Typen von Paketen werden unterschiedlich behandelt, insbesondere wird Verkehr, der dem vorgegebenen Profil genügt, bevorzugt behandelt. Konkret werden in einem solchen Schema die Pakete eines Nutzers, der sich korrekt verhält, als "In" markiert, während Pakete eines Nutzers, der sein Profil überschreitet, eine "Out"-Markierung erhalten. In Überlastsituationen ist es dann ein Leichtes, bevorzugt "Out"-Pakete auszusondern, ohne daß der Verkehr in den Routern irgendwie in Flows oder Warteschlangen aufgeteilt sein muß.

Im Gegensatz hierzu beschäftigt sich [Kil+98] mit der ebenso naheliegenden Idee, nicht die Nutzer, sondern vielmehr die Dienste zu klassifizieren, wobei innerhalb einer Dienstklasse jeder Nutzer gleich zu behandeln ist, während höhere Dienstklassen einen entsprechend höherwertigen Dienst anbieten als niedrigere Klassen und deshalb auch mehr kosten. Dazu wird vorgeschlagen, beispielsweise die Nominal Bit Rate (NBR) als grundlegenden Parameter für eine monatliche Gebühr zu verwenden. Überlastsituationen können durch Monitoring des Auslastungsgrades der Ausgangspuffer in den Knoten erkannt werden; das System reagiert dann durch Aussonderung von Paketen, bevorzugt aus Flows, die ein überhöhtes Verhältnis von momentaner Rate zu NBR aufweisen. Dabei führt jedes Paket Drop-Preference- bzw. Delay-Indication-Bits mit sich, die dem System Anhaltspunkte für den eventuellen Aussonderungsvorgang geben.

Ein weiterer interessanter Vorschlag [Odl97] basiert auf der Aufteilung des Netzwerks in verschiedene logische Unternetzwerke, die zwar jeweils alle Pakete auf Best-Effort-Basis behandeln, sich aber voneinander durch unterschiedlich hohe Preise unterscheiden. Erfahrungen, die man mit einem solchen Vorgehen in der Pariser U-Bahn gesammelt hat (woher der Vorschlag auch seinen Namen PMP, Paris Metro Pricing, erhalten hat [Odl99]) lassen erwarten, daß die teureren Netze weniger häufig frequentiert werden und dadurch einen höherwertigen Dienst anbieten können, ohne dafür jedoch formale Garantien auszusprechen.

## 6.3 Anforderungen aus der Sicht eines Charging-und-Accounting-Tools

Wie aus dem in Abschnitt 6.2 gegebenen Literaturüberblick unschwer hervorgeht, liegt der Schwerpunkt bisher durchgeführter Untersuchungen zur Tarifierung von Internetverkehr vor allem in der Entwicklung und Bewertung theoretischer Preismodelle, wobei die Frage nach ihrer praktischen Brauchbarkeit in vielen Fällen weitgehend ausgeklammert bleibt (vgl. hierzu die detaillierte Untersuchung in [Flö98]). Angesichts dieser Tatsache zielt die Hauptrichtung innerhalb des vom Schweizer Nationalfond (SNF) geförderten Projekts CATI (Charging and Accounting Technology for the Internet[1]) darauf ab, eine lauffähige Plattform zu bauen, die die Gebührenerhebung für zukünftige integrierte Internetdienste ermöglicht [SBGP99]. Da sich das heute dominierende IP-Protokoll hinsichtlich der Paketbehandlung lediglich eines Best-Effort-Ansatzes bedient, der für zukünftige (echtzeitfähige) Anwendungen wie IP-Telefonie oder Video-on-Demand nicht unbedingt geeignet ist, wird hierzu von einem Reservierungsprotokoll ausgegangen, in das im weiteren Verlauf ein passendes Preismodell zu integrieren ist. Als geeigneter Kandidat hierfür hat sich das von [ZDE+93] erstmals vorgeschlagene und im

1. Genaueres zu diesem Projekt wird in Abschnitt 9.3 ausgeführt. An dieser Stelle sei nur soviel bemerkt, daß unter “Accounting” das Sammeln von Daten über Ressourcenkonsum zu verstehen ist, während sich “Charging” auf die Transformation dieser Informationen in monetäre Einheiten (z.B. über Preismodelle) bezieht.

RFC 2205 [BZB+97] detailliert ausgearbeitete RSVP (Resource ReSerVation Protocol) herausgestellt.

Im nächsten Abschnitt folgt daher nach einem kurzen Überblick über die Hauptausrichtung zukünftiger Internet-Technologie, wie sie sich in den IntServ- und DiffServ-Ansätzen zeigt, eine allgemeine Einführung in die Grundideen von RSVP. Im Anschluß daran wird erläutert, wie sich dieses Protokoll grundsätzlich für die Übertragung von Gebühreninformationen nutzbar machen läßt, um schließlich stichwortartig die Anforderungen an eine passendes Preismodell zu skizzieren, die eine Plattform aufwirft, wie sie innerhalb des CATI-Projektes entsteht. Die darauffolgenden Kapitel werden dann zwei Ansätze für ein derartiges Preismodell näher beleuchten.

### 6.3.1 IntServ und DiffServ

Wie im Abschnitt 6.1.2 unter dem Stichwort “Dienstkategorien” bereits kurz angerissen, werden sich in zukünftiger Internet-Technologie zwei Ansätze gegenüberstehen, die gewöhnlich unter den Schlagwörtern “Integrated Services” (IntServ) bzw. “Differentiated Services” (DiffServ) zusammengefaßt werden.

#### Der IntServ-Ansatz

Das Internet in seiner ursprünglichen Form beruht bekanntlich auf einem Best-Effort-Ansatz, demzufolge einzelne Pakete unterschiedslos verworfen werden können, sobald die vorhandenen Ressourcen nicht mehr für einen Weitertransport ausreichen, z.B. falls ein Puffer in einem Switch überläuft. Es wird also keinerlei Garantie für das Ausliefern eines Paketes übernommen. Das mit dem Aufkommen neuartiger Anwendungen wie IP-Telefonie oder Videokonferenzen steigende Bedürfnis nach derartigen Quality-of-Service- (QoS-)Garantien führte zunächst zu einem Ansatz, der dieser Problematik durch eine Erweiterung der Infrastruktur des gegenwärtigen Internets zu begegnen sucht.

Die Leitidee dieser Erweiterung besteht in der Einführung eines “Flow-Konzepts”. Während im herkömmlichen Best-Effort-Betrieb die zu einem Datenstrom gehörenden Einzelpakete unabhängig voneinander durch das Netz geroutet und erst am anderen Ende wieder zusammengefaßt wurden, beruht der IntServ-Ansatz darauf, eine solche Zusammenfassung jeweils von Router zu Router durchzuführen. Die entstehende Gruppierung von Paketen wird als “Flow” bezeichnet. Kernstück der Behandlung von Flows ist die Möglichkeit, über ein entsprechendes Protokoll, wie es exemplarisch in Abschnitt 6.3.2 vorgestellt wird, vorab entsprechende Ressourcen für ihren Ende-zu-Ende-Transport zu reservieren. Dies geschieht durch Einrichtung eines “Per-Flow-States” in den Routern, der die wesentlichen Charakteristiken beinhaltet, nach denen die zum Flow gehörenden Pakete zu verarbeiten sind. Nachdem ein Reservationswunsch aber auch abgelehnt werden kann, ergibt sich hieraus vor allem aber auch die Möglichkeit,

über eine geeignete Zugangskontrolle QoS-Garantien (z.B. hinsichtlich Bandbreite, Verzögerung, Jitter etc.) für einzelne Flows auszusprechen.

Hieraus entstehende Dienstklassen umfassen z.B. den “Guaranteed Service”, der eine absolute Garantie über die vom Netz bereitzustellenden QoS-Parameter beinhaltet und entsprechend harte Anforderungen an die Verkehrskontrolleinheiten stellt. Der “Controlled-Load Service” erwartet von der entsprechenden Anforderung genügend Toleranz, um kurzzeitig mit einer verminderten Übertragungsqualität auszukommen, falls unerwarteterweise eine plötzlichen Schwankung der Verkehrslast (etwa in Form eines Bursts) eintritt. Der herkömmliche “Best-Effort Service” gibt auch im neuen Kontext keinerlei Garantien ab, sondern behandelt alle Pakete unterschiedslos nach dem FIFO-Prinzip. Für nähere Details sei z.B. auf [SBC94] verwiesen.

### Der DiffServ-Ansatz

Nachdem sich herauskristallisierte, daß die Komplexität des IntServ-Ansatzes im Falle von Transitnetzen mit hohen Flowanzahlen stark ansteigt, begann die Suche nach einer global skalierbaren Dienstarchitektur. Der entstehende IETF-Vorschlag [BBC+98] erhielt den Namen “DiffServ” und besteht im Kern aus der Klassifikation einzelner Pakete durch eine entsprechende Markierung im IP-Header, die jeweils beim Betreten des Netzes (also am “Edge”) vorgenommen wird. Dabei wird jeder Paketquelle über einen entsprechenden Vertrag, das “Service Level Agreement” SLA, ein bestimmtes Profil zugewiesen und die Pakete je nachdem markiert, ob sie innerhalb der im SLA festgelegten Grenzen liegen oder nicht. Letztere “Out-of-Profile”-Pakete können dann von Internet Service Providern (ISP) ggf. unterschiedlich behandelt (z. B. in Form von Best Effort) oder auch einfach verworfen werden.

Die entstehende Architektur läßt sich durch lokales Netzverhalten, das sog. “Per-Hop-Behaviour” (PHB), beschreiben, womit das von außen beobachtbare Verhalten von Paketen, die zu einer bestimmten Verkehrsklasse gehören, erfaßt wird. Beispiele für bereits definierte PHBs umfassen das “Expedited Forwarding”, das (in der Art eines “Premium Services”) mit hoher Wahrscheinlichkeit genügend Bandbreite zur Verfügung stellt, “Assured Forwarding” als ein Prioritätsdienst, der auf Verkehrsklassen und Drop-Präferenzen basiert, oder wiederum “Best Effort” als Default-PHB. Daneben ist auch das Aushandeln von SLAs derzeit Gegenstand intensiver Forschungsarbeit. Für weitere Details hierzu verweisen wir auf [FSP99].

### IntServ- über DiffServ-Szenarien

Zusammenfassend halten wir also fest, daß IntServ durch die Möglichkeit, Ressourcen zu reservieren, es erlaubt, harte QoS-Garantien auszusprechen, allerdings um den Preis schlechter Skalierbarkeit in Transit- bzw. Backbone-Netzen mit den hierin auftretenden riesigen Flowzahlen, während DiffServ genau dieses Skalierbarkeitsproblem umgeht, dadurch aber nurmehr weichere QoS-Garantien ermöglicht. Um beide Welten zu vereinen, liegt es nahe, ein IntServ-

over-DiffServ-Szenario zu untersuchen, wie es in Abbildung 6-2 skizziert ist. Hierbei wird zwischen sog. “Core-” und “Access”-Netzwerken unterschieden. Während die Access-Netzwerke dazu dienen, die Einzelnutzer an das Internet anzubinden, bilden die Core-Netzwerke den Backbone des Internets. Entsprechend wird davon ausgegangen, daß die Access-Netzwerke auf IntServ basieren und entsprechende Ende-zu-Ende-Reservationen durchführen, während das Core-Netzwerk DiffServ-orientiert arbeitet. Grundsätzlich werden dabei reservierte Flows durch Aggregationsmechanismen zu noch größeren Gruppen zusammengefaßt, was eine Erhöhung der Dienstqualität aufgrund der entstehenden Multiplexing-Vorteile erlaubt, woraufhin durch geeignetes “Tunneling” die Distanz zwischen dem entsprechenden Ingress- und Egress-Router der “DiffServ-Wolke” überbrückt wird. Auf diese Weise wird insbesondere die Einrichtung von Virtual Private Networks (VPNs) unter Aufsetzen auf das Internet möglich. Nähere Einzelheiten hierzu sind u. a. in [SBGP99] und [BYF+99] beschrieben.

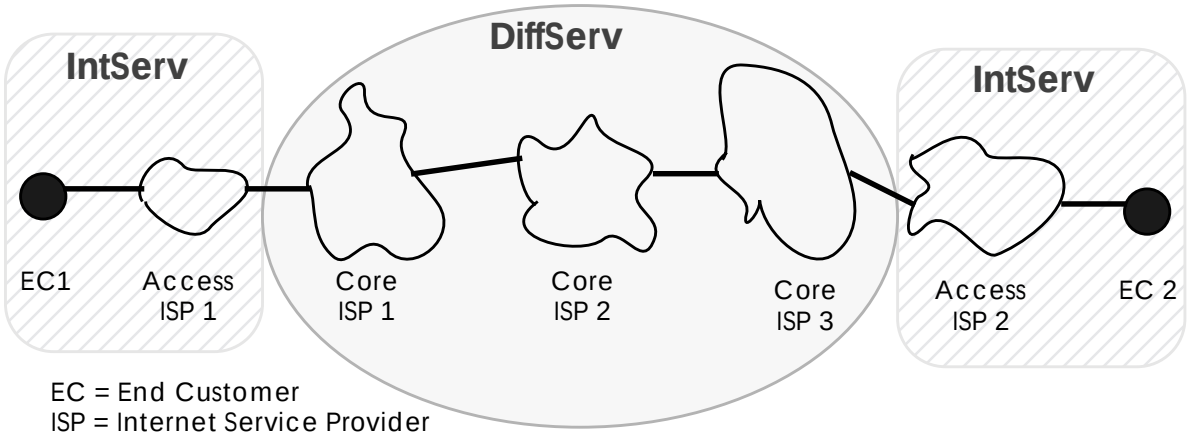


**Abbildung 6-2:** Ein IntServ-over-DiffServ-Szenario

### 6.3.2 RSVP

Wie in [ZDE+93] erläutert wird, hat eine IntServ-basierte Netzarchitektur zumindest fünf Komponenten zu berücksichtigen, nämlich

- die Spezifikation von Flows, die eine Charakterisierung des vom Sender ausgehenden Verkehrsstroms wie der Dienstanforderungen der entsprechenden Anwendungen in einer sog. “flowspec” formuliert (wobei die Zuordnung der Pakete zu den einzelnen Flows durch einen entsprechenden “Classifyer” zustandekommt);
- ein Routing-Protokoll, um über den Pfad zu entscheiden, auf dem die zu einem Flow gehörenden Pakete vom Sender zum Empfänger gelangen;
- ein Reservierungsprotokoll, das die Einführung von QoS-Garantien für Flows dadurch ermöglicht, daß die hierfür benötigten Ressourcen vorab reserviert werden können;

- ein Algorithmus zur Zugangskontrolle, um zu entscheiden, welche Anforderungen angenommen werden können und welche abzuweisen sind, um eingegangene QoS-Garantien erfüllen zu können;
- ein Algorithmus zum Paket-Scheduling, der innerhalb eines Switches jeweils das Paket auswählt, das als nächstes zu übertragen ist.

### Grundidee des Reservierungsprotokolls RSVP

RSVP ist ein schon sehr weit ausgereifter und auch bereits implementierter Vorschlag für ein Reservierungsprotokoll. Es wird grundsätzlich vom Sender eines Flows initiiert, der eine *PATH*-Message zum Empfänger schickt. Diese enthält u.a. die Spezifikation des Flows (beispielsweise die benötigte Bandbreite) und wird mittels Standardprotokollen zum Empfänger geroutet. Auf dem Pfad der Nachricht merkt sich jeder Knoten den vorhergehenden Hop (also den Knoten, von dem er die Nachricht erhalten hat). Dadurch wird es möglich, die vom Empfänger als nächstes verschickte *RESV*-Message, die erst die tatsächliche Reservation z.B. der benötigten Bandbreite in den Knoten auslöst, auf demselben Pfad zurückzuschicken, auf dem die PATH-Message gekommen ist.

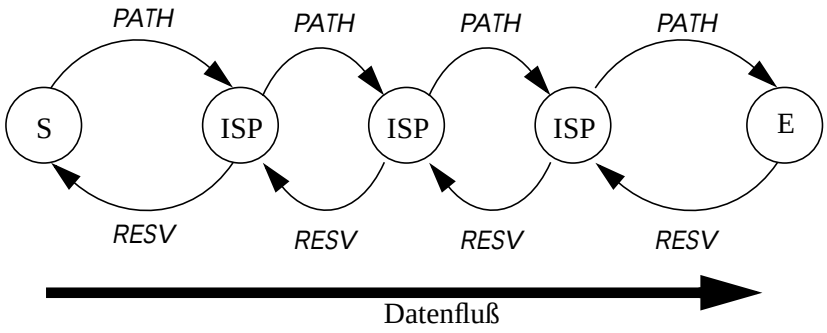


**Abbildung 6-3:** PATH-Message und RESV-Message vom Sender (S) über Provider (ISP) zum Empfänger (E) und schließlich zurück zum Sender (S) entlang eines Verbindungspfads

Daß die Reservierung erst aufgrund der vom Empfänger verschickten RESV-Nachricht durchgeführt wird, erlaubt die Abstimmung der für den Empfänger möglichen Quality-of-Service mit der vom Sender vorgeschlagenen, was insbesondere im Fall von Multicast-Übertragungen an verschieden gut ausgerüstete Empfänger von Vorteil ist. Abbildung 6-3 zeigt, wie der Reservierungsmechanismus auf der Senderseite der Verbindung startet und vom Empfänger beantwortet wird.

Neben den eingeführten gibt es unter RSVP noch weitere Nachrichtentypen, insbesondere *ERROR*-Messages zur Fehlermeldung sowie *PATH-TEARDOWN* bzw. *RESV-TEARDOWN*-Messages zur regulären Beendigung von Reservationen.

Der allgemeine Aufbau einer RSVP-Nachricht ist in Abbildung 6-4 dargestellt. Demnach erfolgt der Transport der Nachrichten im Payload gewöhnlicher IP-Datagramme, wobei auf

den RSVP-Header eine Sammlung von sog. RSVP-Objekten folgt, die den “Body” der RSVP-Nachricht ausmachen.

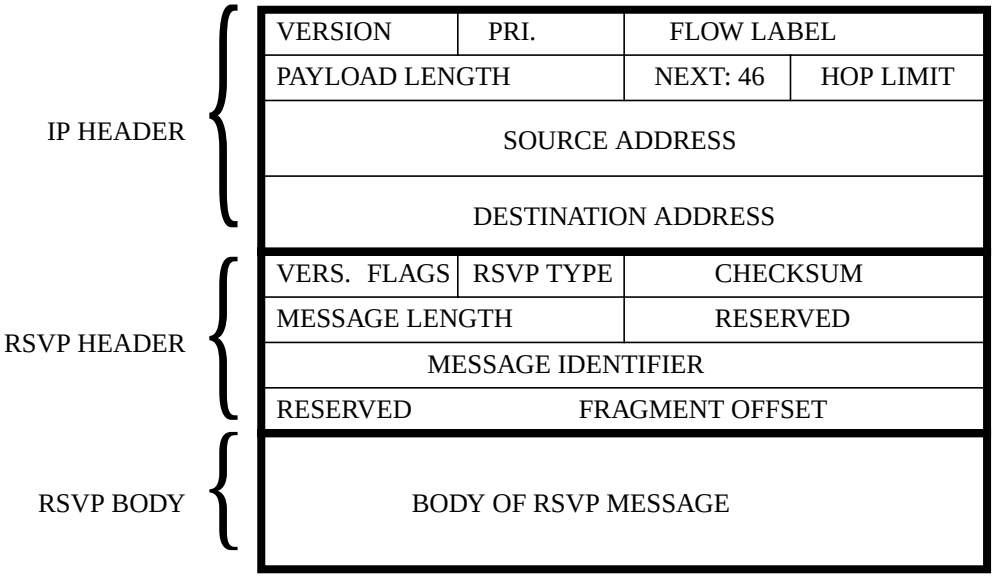


**Abbildung 6-4:** Format einer RSVP-Nachricht (nach RFC 2205 [BZB+97]).

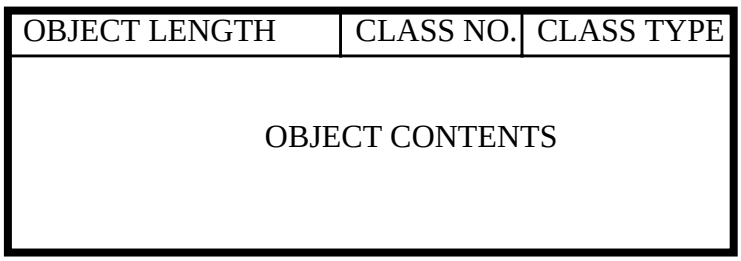


**Abbildung 6-5:** Format eines RSVP-Objekts (nach RFC 2205 [BZB+97])

Abbildung 6-5 stellt das Grundformat dar, nach dem sämtliche RSVP-Objekte aufgebaut sind. Momentan sind im Standard-RSVP 15 verschiedene Objekte definiert (für nähere Details sei auf [Fos99] verwiesen).

### Ein Charging-Framework für RSVP

Aufbauend auf dieser allgemeinen Protokollspezifikation wird in [FSVP98] (in ähnlicher Form auch in [KSWS98]) untersucht, wie sich die beschriebenen Funktionalitäten von RSVP für den Austausch von Gebühren- und Abrechnungsinformationen anpassen lassen. Ein wichtiger Gesichtspunkt hierbei ist, daß sich die übertragene Information je nach unterliegendem ökonomischen Modell unterschiedlich interpretieren lassen muß. Die grundlegende Idee besteht in der Nutzung der PATH- und RESV-Nachrichten für die Übertragung von Preisinformationen.

Erweiterte PATH-Nachrichten weisen ein Feld dafür auf, das anfänglich zu Null initialisiert wird. An jedem Hop mit einem Ausgangslink wird der aktuelle Marktpreis für die angeforderte QoS zum Preisfeld hinzuaddiert. Wenn die PATH-Nachricht schließlich beim Empfänger angekommen ist, kann sie zumindest näherungsweise ein Bild von der momentanen Marktsituation vermitteln, auch wenn kurzfristige Schwankungen in der Netzauslastung dazu führen können, daß der wirkliche Preis schlußendlich noch davon abweicht. Über die zurückgesandte RESV-Nachricht wird dann auch dem Sender der derzeitige Preis für die Verbindung mitgeteilt. Darüberhinaus lassen sich diese Nachrichten auch für die Abrechnung selbst verwenden, beispielsweise können PATH-Nachrichten Zahlungen des Senders beinhalten, während die Zahlungen des Empfängers über RESV-Nachrichten abgewickelt werden.

### 6.3.3 Preismodelle in Multiprovider-Szenarien

Vorliegendes Kapitel diente einer ersten Einführung in Kontext und Problematik der Modellierung von Internet-Tarifen. Nach der Vorstellung eines Klassifikationsschemas für Preismodelle wurde ein Überblick über Related Work in diesem Bereich gegeben, bevor anhand der Erfordernisse des CATI-Projektes detaillierter auf einige der grundlegenden Konzepte eingegangen wurde, die im Umfeld der praktischen Realisierung eines Charging-and-Accounting-Tools von Bedeutung werden. Abbildung 6-6 faßt stichpunktartig die wichtigsten der sich hieraus ergebenden Anforderungen an ein geeignetes Preismodell zusammen.

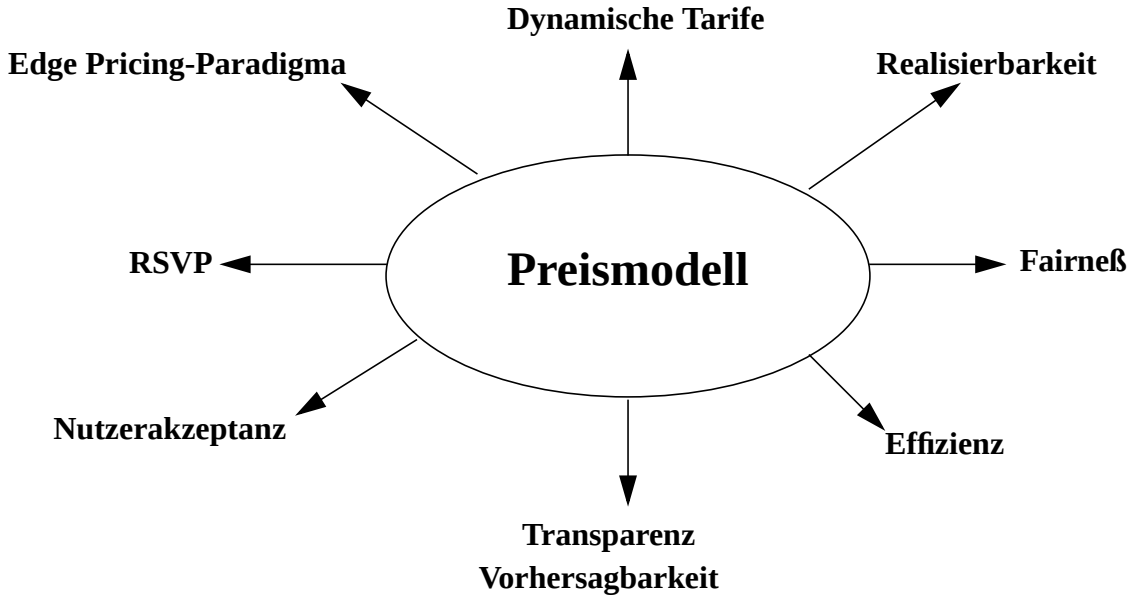


**Abbildung 6-6:** Anforderungen an ein Preismodell für Integrierte Internet-Dienste

Neben der Verwendung von RSVP als Reservierungsprotokoll ist hierbei insbesondere die Multiprovidersicht von zentraler Bedeutung. Die einzelnen ISPs werden hierbei im wesentlichen als Knoten modelliert, die lokal den Preis für die Benutzung der entsprechenden Netzwerkressourcen festsetzen. Anders ausgedrückt besitzt jeder ISP eine Art “Black Box”, in die er je nach gusto ein ihm angenehmes Preismodell stecken kann, um damit Gebühren für die Nutzung seines Netzes festzulegen.

Die Preisermittlung selbst kann auf zwei unterschiedliche Arten geschehen. Entweder geht man von einer Konkurrenzsituation unter den ISPs aus, die in diesem Fall aus den ihnen zugänglichen Informationen (z.B. der aktuellen Auslastung bestimmter Verbindungen) auf geschickte Weise einen Preis für die Annahme einer neuen Verbindung bestimmen müssen. Oder aber man läßt die Nutzer um die Ressourcen konkurrieren, beispielsweise in Form einer Auktion.

Die folgenden Kapitel beschäftigen sich mit diesen beiden Ansätzen. Zunächst wird in Kapitel 7 ein ursprünglich aus der Telefonie stammendes Modell der lokalen Preisbestimmung für eine Verbindung anhand ihrer aktuellen Auslastung auf die im Internet herrschenden Verhältnisse angepaßt. Kapitel 8 beschäftigt sich dann im hauptsächlich mit Modellen, die verstärkt auf die Sicht der jeweiligen Nutzer zugeschnitten sind, insbesondere mit der Untersuchung eines neuen Auktionsschemas für Multiproviderszenarien.

KAPITEL 7

# *Verallgemeinerte Preisfunktionen in Stochastischen Verlustnetzen*

## 7.1 Modell und Preisfunktionen

Dieses Kapitel ist der Frage gewidmet, wie sich der Preis für eine Ressource (z. B. ein Internetlink) aus lokal vorliegenden Informationen über ihre momentane Auslastung ableiten läßt. Der hier beschriebene und weiterentwickelte Vorschlag geht auf ein bislang nicht weiter beachtetes Nebenprodukt eines Ansatzes von Kelly zurück, der in [Kel94] eine untere Schranke für die Verlustrate in vollvermaschten Telefonnetzen beschreibt. Seine Idee wurde in einer Reihe nachfolgender Arbeiten aufgegriffen und vertieft (u.a. in [Rei92], [GR93], [Rei94], [GR95] und [GK95]) und hat in der Zwischenzeit unter dem Namen “Kelly’s Bound” sogar Einzug in die Lehrbuchliteratur gehalten [Ros95].

### 7.1.1 Kelly’s Bound

Der ursprüngliche Kontext, der zur Formulierung von Kelly’s Bound führte, läßt sich summarisch wie folgt skizzieren (für das allgemein zugrundeliegende Modell vgl. [Kel79] und spezieller [Kel91b]): Gegeben sei ein vollvermaschtes stochastisches Verlustnetzwerk (etwa ein Telefonnetz), bei dem zwischen jedem Knotenpaar $d \in D$ eine direkte Verbindung der Kapazität $C \geq 0$ Kanäle besteht. Gesprächswünsche kommen für jedes Knotenpaar $d \in D$ als Poissonprozeß mit Rate $\nu_d$ an; sie werden entweder akzeptiert und belegen dann für eine exponentialverteilte Verweilzeit mit Mittelwert 1 genau einen Kanal, oder sie werden blockiert und gehen dann verloren. Kann ein Gespräch nicht auf der direkten Verbindung $i(d) \in I$ zwischen Sender- und Empfängerknoten geroutet werden, dann besteht die Möglichkeit, aus einem entsprechenden Pool $R(d)$ eine indirekte Route $r \in R(d)$ (via einen dritten Knoten) zu wählen. Bezeichnet nun $x_i$ den direkten Verkehrfluß auf Verbindung $i = i(d)$ und $y_r$ den indirekten Verkehrsfluß auf Route $r$, so führt die Frage nach dem unter diesen Voraussetzungen maximal möglichen Durchsatz des Gesamtsystems auf folgendes Optimierungsproblem [Rei94]:

$$max \sum_{d \in D} \left( x_{i(d)} + \sum_{r \in R(d)} y_r \right) \text{ über } x_i \geq 0, y_r \geq 0 \tag{7.1}$$

unter den Nebenbedingungen

$$x_{i(d)} + \sum_{r \in R(d)} y_r \leq \nu_d \qquad \forall d \in D \tag{7.2}$$

$$\sum_{d \in D} \sum_{r \in R(d):\, i \in r} y_r \leq M_i(x_i) \qquad \forall i \in I \tag{7.3}$$

Hierbei verhindert (7.2), daß mehr Gespräche stattfinden als angefragt werden, und (7.3) beschreibt die Tatsache, daß nur soviele indirekte Gespräche auf einer bestimmten Verbindung stattfinden dürfen, daß der Fluß $x_i$ der direkten Gespräche davon nicht beeinträchtigt wird; entsprechend beschreibt die Funktion $M_i(x_i)$ die maximale Anzahl der auf $i$ zulässigen indirekten Gespräche, falls schon $x_i$ direkte Gespräche über diese Verbindung laufen.

Die zentrale Idee von Kelly's Bound liegt dabei in der Brechung des komplexen Gesamtproblems herunter auf die Ebene der einzelnen Verbindungen, und hier wiederum in der Bestimmung der erläuterten Funktion $M_i(x_i)$. Die Berechnung dieser Funktion, auf deren Details wir im nächsten Abschnitt zurückkommen werden, erlaubt es dann, für jede Verbindung des Netzes eine Nebenbedingung der Form (7.3) zu formulieren, die sich wiederum als eine Sammlung linearer Gleichungen darstellen läßt (vgl. [Rei94]) und damit eine effiziente Lösung von (7.1) als lineares Optimierungsproblem ermöglicht.

Soweit eine kurze Einführung in den grundsätzliche Methodik von Kelly's Bound. Unser weiteres Vorgehen wird sich auf die Untersuchung einer einzelnen Verbindung beschränken und hierfür nach einer genauen Darstellung der zugrundeliegenden Modellannahmen zeigen, wie sich Kelly's Ansatz für die Ableitung einer Funktion nutzen läßt, die es erlaubt, in Abhängigkeit von der Auslastung einer Verbindung einen Preis für die Annahme eines neu ankommenden Gesprächs anzugeben.

### 7.1.2 Modellbeschreibung

Das im folgenden für unsere Zwecke verwendete Modell ist etwas allgemeiner als das eben beschriebene, insbesondere geht es einfach von zwei unterschiedlichen Klassen von Gesprächen (mit in gewissem Sinne unterschiedlichem "Wert") aus, läßt aber zunächst offen, in welchem detaillierten Sinn die erste Klasse gegenüber der zweiten als priorisiert angesehen wird. Wir betrachten also eine isolierte Ressource mit Kapazität $C$ Kanälen, die von zwei Klassen von Gesprächen benutzt wird: A-Gespräche kommen stochastisch als Poissonprozeß der Rate $\nu$ an, während B-Gespräche zu jedem Zeitpunkt in beliebiger Menge zur Verfügung stehen, d. h. einen "deterministischen Ankunftsprozeß" der Rate $\infty$ aufweisen (was im Falle des

ursprünglichen Modells mit direkten und indirekten Gesprächen eine durchaus realistische Annahme ist, wie [MGH91] gezeigt hat). Die Verweilzeiten von Gesprächen beider Klassen sind exponentialverteilt mit Mittelwert 1. Akzeptierte Gespräche erhalten für die Dauer ihrer Verweilzeit einen Kanal zugeteilt, nicht akzeptierte Gespräche gehen verloren. Ferner bezeichne $M(x)$ den maximal möglichen Fluß an B-Gesprächen, wenn mindestens ein Fluß von $x$ A-Gesprächen über die Verbindung läuft.

Für die weitere Behandlung dieses Modells entscheidend ist die Tatsache, daß das für die Maximierung des Durchsatzes optimale Vorgehen bekannt ist [Key90]. Es handelt sich dabei nämlich stets um eine *Trunk Reservation Policy*, die sich folgendermaßen beschreiben läßt (vgl. Abbildung 7-1): Zunächst wird ein ganzzahliger Parameter $a$, der *Trunk Reservation Parameter*, spezifiziert. Die Zulassung von Gesprächen hängt dann von der momentanen Auslastung der Verbindung ab: sind gegenwärtig weniger als $C - a$ Kanäle belegt, werden ankommende A- wie B-Gespräche akzeptiert, steigt die Auslastung aber über diese Schranke, so kommt nur noch die Zulassung von A-Gesprächen in Frage.

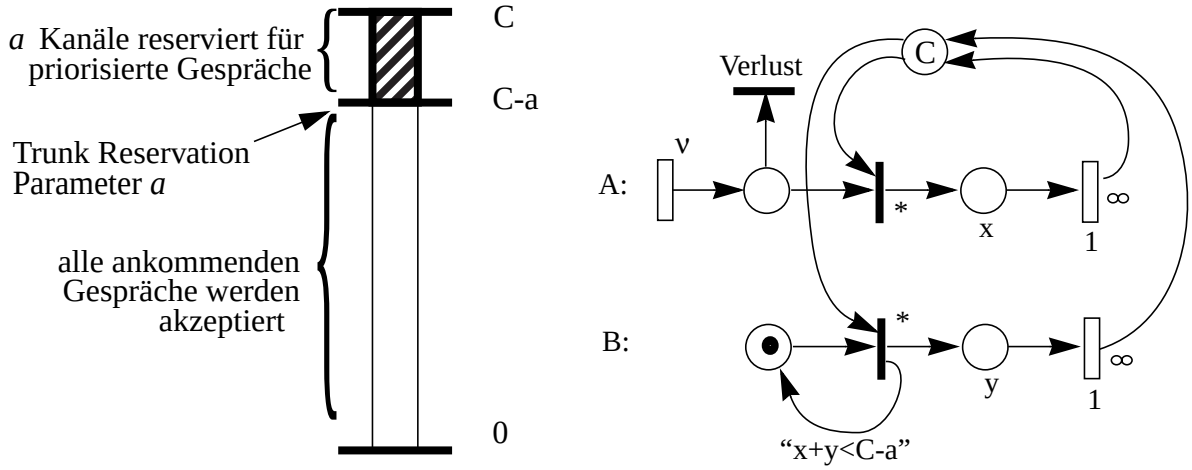


**Abbildung 7-1:** Zur Strategie der Trunk Reservation (rechts in Petrinetz-Darstellung)

Nun ist es möglich, unter Benutzung dieser (optimalen) Trunk-Reservation-Strategie eine über den Reservierungsparameter $a$ parametrisierte Darstellung der Funktion $M(x)$ zu gewinnen. Betrachten wir also eine einzelne Verbindung, die die Zulassung von neu ankommenden A- bzw. B-Gesprächen anhand des Trunk Reservation-Parameters $a$ trifft. Der Zustand diese Verbindung läßt sich dann in Form eines Geburts- und Todesprozesses angeben, vgl. Abbildung 7-2. $\pi_a(C-a)$ beschreibt hierbei die stationäre Wahrscheinlichkeit, daß sich der durch den Parameter $a$ charakterisierte Prozeß in Zustand $C - a$ aufhält.

Wir sehen, daß unter Trunk Reservation ein B-Gespräch sofort akzeptiert wird, wenn der Zustand der Ressource von $C - a$ auf $C - a - 1$ herabspringt. Die Rate, mit der dies geschieht,

ergibt sich als Produkt aus der Wahrscheinlichkeit, daß sich das System im Zustand $C - a$ befindet, und der Rate, mit der in diesem Zustand Gespräche beendet werden (da jedes der $C - a$ Gespräche im Schnitt eine Gesprächsdauer von 1 aufweist, ist letztere gleich $C - a$):

$$(C - a) \cdot \pi_a(C - a) = M_a . \tag{7.4}$$

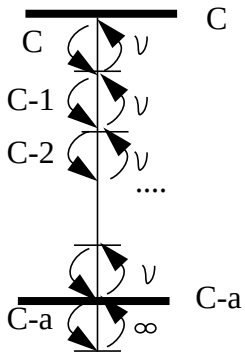


**Abbildung 7-2:** Beschreibung der Ressourcenauslastung als Geburts- und Todesprozeß

Während (7.4) also die Rate der akzeptierten B-Gespräche angibt, ergibt sich der entsprechende Fluß an A-Gesprächen zu

$$x_a = \nu \cdot (1 - \pi_a(C)), \tag{7.5}$$

weil ein mit Rate $\nu$ ankommendes A-Gespräch mit Wahrscheinlichkeit $(1 - \pi_a(C))$ noch einen freien Kanal in der Ressource vorfindet und deshalb akzeptiert werden kann.

.

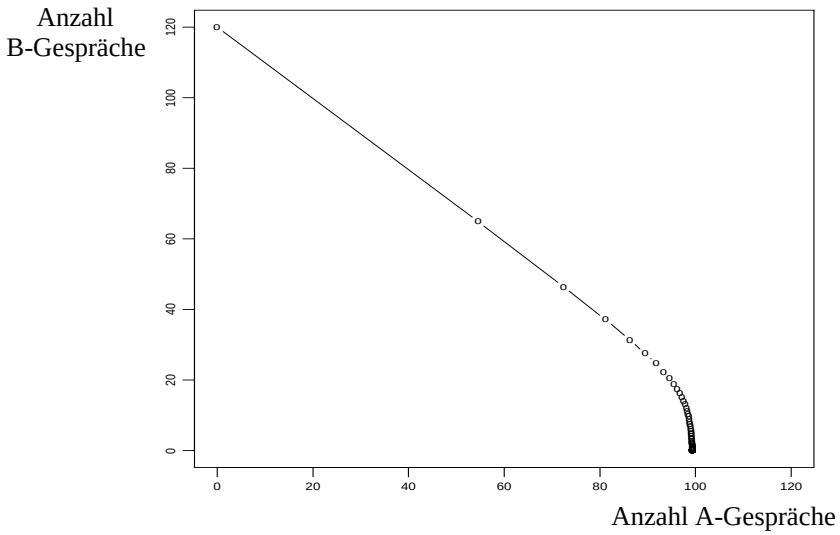


**Abbildung 7-3:** Die Funktion *M* maximal möglicher B-Gespräche (für $C = 120$ und $\nu = 100$)

Auf diese Weise läßt sich für jeden Trunk Reservation-Parameter $a \in \{0, 1, \ldots, C\}$ der entsprechende Fluß von A- und B-Gesprächen berechnen. Die Optimalitätseigenschaft der Trunk Reservation Policy hat dabei zur Folge, daß (7.4) in Wirklichkeit den maximal möglichen Fluß an B-Gesprächen angibt, wenn der Fluß an A-Gesprächen $x_a$ gem. (7.5) beträgt. Läßt man also den Parameter $a \in \{0, 1, \ldots, C\}$ laufen, so ergibt sich schließlich die in Abbildung 7-3 dargestellte Parametrisierung der Funktion $M(x)$ (vgl. [Rei94]).

### 7.1.3 Ableitung der Preisfunktion

Jetzt wollen wir das Modell allerdings von einer etwas anderen Seite her unter die Lupe nehmen [Rei99b]: Angenommen, ein A-Gespräch zahlt im Fall des Zustandekommens einen Betrag von 1 an die Verbindung, wohingegen ein B-Gespräch, das ja "weniger wert" ist, einen niedrigeren Betrag entrichtet. Ferner seien im Augenblick $m$ Kanäle der Ressource durch laufende Gespräche belegt. Wie kann man herausfinden, welchen Betrag ein B-Gespräch, das ja sofort zur Verfügung steht, entrichten muß, damit es sich für den Betreiber der Ressource auf lange Sicht nicht lohnt, das B-Gespräch nicht zu akzeptieren und stattdessen auf ein neues A-Gespräch zu warten, das aber dafür mehr einbringt?

Wir wissen, daß Trunk Reservation nach wie vor die optimale Politik für die Maximierung der Einnahmen aus der Verbindung darstellt. Sei also $R(a;p) = x(a) + p \cdot M(x(a))$ definiert als die Einnahme für den Betreiber, falls die Verbindung mit dem Trunk Reservation-Parameter $a$ operiert und A-Gespräche einen Betrag von 1 entrichten, während B-Gespräche den Betrag $p$ zahlen; $x(a)$ und $M(x(a)) = M_a$ sind dabei die Gesprächsflüsse gemäß (7.4) bzw. (7.5) [Rei99a]. In Abhängigkeit vom Trunk Reservation-Parameter $a$ ist dann die Funktion $R$ maximal für

$$\frac{d}{da}R(a;p) = 0 \tag{7.6}$$

Im Blick auf unsere ursprüngliche Frage sei nun $p$ definiert als derjenige Preis für ein B-Gespräch, bei dem es (bei momentaner Verbindungsauslastung von $m$) für die Gesamteinnahme aus der Verbindung keine Rolle spielt, ob der Betreiber sofort ein B-Gespräch akzeptiert oder stattdessen auf ein neues A-Gespräch wartet. Dieser Sachverhalt läßt sich aber auch als Wahl zwischen zwei benachbarten Trunk Reservation-Parametern darstellen. Sei hierzu $\hat{a} = C - m - 1$. Operiert nun die Verbindung mit dem Parameter $\hat{a}$, so ist dies gleichbedeutend damit, daß die momentane Auslastung der Verbindung die Zulassung eines weiteren B-Gesprächs erlaubt (vgl. Abbildung 7-1). Entscheidet man sich dagegen für den Trunk Reservation-Parameter $\hat{a} + 1 = C - m$, so entspricht dies einer Ablehnung des B-Gesprächs zugunsten des Wartens auf ein neues A-Gespräch.

Damit ergibt (7.6) aufgrund der diskreten Natur des Trunk Reservation-Parameters

$$\frac{d}{d\hat{a}}R(\hat{a};p) = \frac{(x_{\hat{a}} + pM_{\hat{a}}) - (x_{\hat{a}+1} + pM_{\hat{a}+1})}{(\hat{a}+1) - \hat{a}} = 0 \tag{7.7}$$

und führt damit schließlich zu folgender sog. “Reward Balance Equation” [Rei99b]

$$x(\hat{a}) + p \cdot M(x(\hat{a})) = x(\hat{a}+1) + p \cdot M(x(\hat{a}+1)) \tag{7.8}$$

wobei $x(\hat{a})$ den Fluß an A-Gesprächen für einen Trunk Reservation-Parameter von $\hat{a}$ beschreibt und $M$ wie oben definiert ist.

Somit erhalten wir als korrekten Preis $p$ für ein neu ankommendes B-Gespräch

$$p = \frac{x(\hat{a}+1) - x(\hat{a})}{M(x(\hat{a})) - M(x(\hat{a}+1))}. \tag{7.9}$$

Löst man den entsprechenden Geburts- und Todesprozeß (vgl. Abbildung 7-2) standardmäßig über seine lokalen Flußgleichungen, so läßt sich (7.9) sogar in geschlossener Form angeben [Rei94] als

$$p(m,\nu) = \nu\left[(\nu - m)\sum_{i=0}^{C-m-1}\frac{C!}{\nu^i(C-i)!} + \frac{C!}{\nu^{C-m-1}m!}\right]^{-1} \tag{7.10}$$

wobei wie oben $C$ die Kapazität der Resource, $m$ ihre derzeitige Auslastung und $\nu$ die Ankunftsrate der A-Gespräche beschreibt.

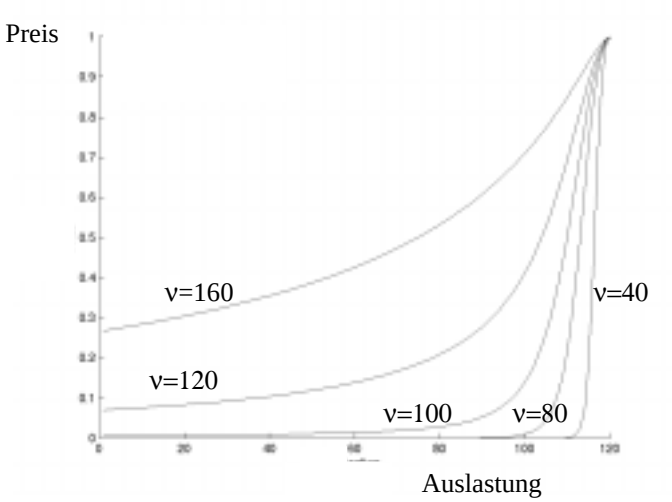


**Abbildung 7-4:** Preisfunktionen für zwei Gesprächsklassen. Die Ressourcenkapazität beträgt $C = 120$, die Poisson-Ankunftsraten der A-Gespräche variieren über $\nu \in \{40, 80, 100, 120, 160\}$.

Abbildung 7-4 veranschaulicht den Verlauf des ermittelten Preises für B-Gespräche in Abhängigkeit von der Auslastung für eine Ressource mit Kapazität $C = 120$ und Ankunftsraten $\nu \in \{40, 80, 100, 120, 160\}$. Wir sehen deutlich, daß ein signifikanter Preis nur bei relativ hoher Auslastung zustande kommt, ferner ist die Preiskurve umso höher, je größer die Konkurrenz durch ankommende A-Gespräche ist (wie man ja auch erwarten würde).

### 7.1.4 Einschränkungen und Erweiterungsansätze

Wie eingangs erwähnt, entstammt der vorgestellte Ansatz ursprünglich der Welt der Telefonie. Der Versuch, ihn für die Preisbestimmung für integrierte Internet-Dienste unmittelbar nutzbar zu machen, stößt sehr schnell auf folgende drei Einschränkungen:

- **Skalierbarkeit:** Die explizite Lösung (7.10) läßt sich zwar für $C = 120$ durchaus in sinnvoller Rechenzeit lösen, erweist sich aber hinsichtlich heute aktueller Bandbreiten (die 622 Mbps einer STM-12-Leitung entsprechen beispielsweise über 35 000 Telefongesprächen) als extrem schlecht skalierbar (was beispielsweise ein Blick auf die notwendigen Fakultätsberechnungen klarstellt).
- **Mehrklassenverkehr:** Läßt man weiterhin mehrere Verkehrsklassen zu, die sich zudem auch noch in ihren wesentlichen Charakteristiken (z.B. ihren Bandbreitenanforderungen oder Gesprächsdauern) unterscheiden können, so führt dies sofort zum Zusammenbruch des Geburts- und Todesprozesses von Abbildung 7-2: Unterschiedliche Bandbreitenanforderungen bedeuten nämlich eine Fülle neu hinzukommender Übergänge, die eine Lösung über die lokalen Flußgleichungen unmöglich machen.
- **Optimale Policy:** Bei Zulassung von mehr als zwei Klassen ist außerdem bis heute kein Pendant zur Trunk Reservation Policy bekannt (vgl. [Ros95], [DS99]), das einen optimalen Durchsatz (bzw. maximale Einnahmen für den Betreiber) garantieren könnte.

Der Rest dieses Kapitels stellt Ansätze zur Lösung dieser drei Fragestellungen vor. Im Zentrum steht dabei die Anwendung eines in etwas anderem Zusammenhang entwickelten Approximationsverfahrens für die stationären Wahrscheinlichkeiten des von uns zu lösenden Markov-Prozesses für Vielklassenverkehr, das im folgenden Abschnitt 7.2 vorgestellt wird. Diese Approximation behebt aufgrund ihrer günstigen Komplexitätseigenschaften sofort das Problem der Skalierbarkeit und ist außerdem explizit für eine Anwendung auf den Fall beliebig vieler unterschiedlicher Verkehrsklassen geeignet, wie die in Abschnitt 7.3 vorgenommene Validierung aufzeigt. In Abschnitt 7.4 schließlich wird ein Ansatz vorgestellt, der das Problem der optimalen Politik für den Mehrklassenfall dadurch behandelt, daß er das Mehrklassenproblem in geeigneter Weise auf eine Folge einzelner Zweiklassenprobleme zurückführt, die ihrerseits mit dem in Abschnitt 7.3 vorgestellten Instrumentarium lösbar sind. Abschnitt 7.5 faßt die Hauptergebnisse dieses Kapitels nochmals kurz zusammen.

## 7.2 Approximation mit UAA und RUAA

Im Kontext von Design und Optimierung von Netzrouting, insbesondere für Virtual Private Networks (VPNs), haben Mitra et al. eine elegante Methode zur näherungsweisen Berechnung der stationären Wahrscheinlichkeitsverteilung $\pi$ der Auslastung einer isolierten Ressource mit Vielklassenverkehr und hoher Bandbreite entwickelt und verfeinert (vgl. u.a. [MM94], [MMR96], [MRM98], [MMR99]). In diesem Abschnitt wird zunächst die Grundidee ihrer Uniform Asymptotic Approximation (UAA) bzw. der daraus hervorgegangenen Refined Uniform Asymptotic Approximation (RUAA) dargestellt, bevor in Abschnitt 7.3 die Näherung durch Vergleich mit den in Abschnitt 7.1 angegebenen exakten Resultaten (vgl. insbesondere Abbildung 7-4) validiert wird.

### 7.2.1 Zur Darstellung der Partitionsfunktion als Kreisintegral

Wir betrachten im folgenden eine Resource der Kapazität $C$, auf der $S$ unterschiedliche Klassen von Verkehr übertragen werden. Gespräche der Klasse $s$ benötigen dabei jeweils $d_s$ Kanäle, kommen unabhängig voneinander als Poissonstrom mit Rate $\lambda_s$ an und dauern im Schnitt $1/\mu_s$, wobei die exponentialverteilten Verweilzeiten (Gesprächsdauern) ebenfalls unabhängig voneinander und von den Ankünften angenommen werden. Aufgrund der sog. Insensitivitätseigenschaft (*Insensitivity Property*) genügt es [MMR99], für jede Klasse $s$ statt Ankunftsrate $\lambda_s$ und Abgangsrate $\mu_s$ deren Quotienten, die *Verkehrsintensität* $\nu_s = \lambda_s/\mu_s$, zu betrachten.

Bezeichne $\pi_{C-j}$ die stationäre Wahrscheinlichkeit dafür, daß die Gesamtanzahl von belegten Kanälen auf unserer Ressource gerade $C-j$ beträgt; wird es später nötig, diese Wahrscheinlichkeit explizit auch noch von der Kapazität $C$ der Ressource abhängig zu machen, so schreiben wir dafür $\pi_{C-j}(C)$. Für diese Wahrscheinlichkeitsverteilung läßt sich zunächst eine einfache Produktformdarstellung herleiten. Hierzu bezeichne $K_s$ die Anzahl an Quellen, die Verkehr der Klasse $s$ produzieren, und $p_s$ die Wahrscheinlichkeit, daß eine solche Quelle unabhängig von den anderen sendet. Dann ist die stationäre Wahrscheinlichkeit, dafür, daß momentan $\boldsymbol{i} = (i_1, i_2, \ldots, i_S)$ Quellen senden, gleich dem Produkt

$$\pi(\boldsymbol{i}) = \frac{1}{G(\boldsymbol{K}, C)} \prod_{s=1}^{S} \binom{K_s}{i_s} p_s^{i_s} (1-p_s)^{K_s - i_s}, \tag{7.11}$$

mit $\boldsymbol{K} = (K_1, K_2, \ldots, K_S)$, wobei jede einzelne Quellenklasse binomial verteilt ist und die Normalisierungskonstante oder *Partitionsfunktion* durch

$$G(\boldsymbol{K}, C) = \sum_{\boldsymbol{i}'\boldsymbol{d} \le C} \left( \prod_{s=1}^{S} \binom{K_s}{i_s} p_s^{i_s} (1-p_s)^{K_s - i_s} \right) \tag{7.12}$$

definiert ist (mit $\boldsymbol{d} = (d_1, d_2, \ldots, d_S)$ als Vektor des Bandbreitenbedarfs der einzelnen Klassen).

Wir werden später anhand von (7.21) sehen, wie sich die stationäre Wahrscheinlichkeit $\pi_{C-j}$ mit Hilfe geeigneter Partitionsfunktionen ausdrücken läßt. *Das bedeutet, daß ein effizientes Verfahren zur Berechnung der Partitionsfunktionen den Schlüssel zur Bestimmung der stationären Wahrscheinlichkeiten bildet.* Deshalb werden wir uns zunächst in Anlehnung an [MM94] um die Berechnung von (7.12) kümmern.

Ausgangspunkt ist die Verwendung der Binomialformel zur Berechnung von

$$(1 - p_s + p_s z^{d_s})^{K_s} = \sum_{i_s=0}^{K_s} \binom{K_s}{i_s} (1-p_s)^{K_s - i_s} (p_s z^{d_s})^{i_s} = \sum_{i_s=0}^{K_s} z^{d_s i_s} \binom{K_s}{i_s} (1-p_s)^{K_s - i_s} p_s^{\,i_s} \quad (7.13)$$

Hieraus erhält man durch Produktbildung

$$\prod_{s=1}^{S} (1 - p_s + p_s z^{d_s})^{K_s} = \prod_{s=1}^{S} \left[ \sum_{i_s=0}^{K_s} z^{d_s i_s} \binom{K_s}{i_s} (1-p_s)^{K_s - i_s} p_s^{\,i_s} \right] \quad (7.14)$$

Die rechte Seite von (7.14) stellt im wesentlichen ein Polynom in der komplexen Variablen $z$ dar, das man nun exponentenweise nach $z^{\boldsymbol{i}'\boldsymbol{d}}$ ordnen kann (hier wie auch schon in (7.12) bezeichnet $\boldsymbol{i}'\boldsymbol{d}$ das übliche Skalarprodukt der beiden Vektoren):

$$\prod_{s=1}^{S} (1 - p_s + p_s z^{d_s})^{K_s} = \sum_{k=0}^{\infty} z^k \left[ \sum_{\boldsymbol{i}'\boldsymbol{d}=k} \left( \prod_{s=1}^{S} \binom{K_s}{i_s} (1-p_s)^{K_s - i_s} p_s^{\,i_s} \right) \right] \quad (7.15)$$

Verwendet man nun noch die bekannte Taylorentwicklung

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n, \quad (7.16)$$

so erhält man aus (7.15)

$$\frac{\prod_{s=1}^{S} (1 - p_s + p_s z^{d_s})^{K_s}}{1-z} = \left( \sum_{n=0}^{\infty} z^n \right) \cdot \left[ \sum_{k=0}^{\infty} z^k \left\{ \sum_{\boldsymbol{i}'\boldsymbol{d}=k} \left( \prod_{s=1}^{S} \binom{K_s}{i_s} (1-p_s)^{K_s - i_s} p_s^{\,i_s} \right) \right\} \right]. \quad (7.17)$$

Ausmultiplizieren von $(1 + z + z^2 + \ldots) \left( \left\{ \sum_{\boldsymbol{i}'\boldsymbol{d}=0} \ldots \right\} + \left( \sum_{\boldsymbol{i}'\boldsymbol{d}=1} \ldots \right) z + \left\{ \sum_{\boldsymbol{i}'\boldsymbol{d}=2} \ldots \right\} z^2 + \ldots \right)$

(rechte Seite) ergibt schließlich mit (7.12)

$$\frac{\prod_{s=1}^{S}(1-p_s+p_s z^{d_s})^{K_s}}{1-z} = \sum_{n=0}^{\infty} z^n \left( \sum_{i'd=0}^{n} \left( \prod_{s=1}^{S} \binom{K_s}{i_s} (1-p_s)^{K_s-i_s} p_s^{i_s} \right) \right) = \sum_{n=0}^{\infty} z^n G(K,n) \tag{7.18}$$

Diese Identität erlaubt uns nun die folgende Rechnung:

$$\frac{1}{2\pi i}\oint \left( \frac{\prod_{s=1}^{S}(1-p_s+p_s z^{d_s})^{K_s}}{1-z} \cdot \frac{1}{z^{C+1}} \right) dz \overset{(7.18)}{=} \frac{1}{2\pi i}\oint \left( \frac{\sum_{n=0}^{\infty} z^n G(K,n)}{z^{C+1}} \right) dz$$
$$= \frac{1}{2\pi i}\oint \left( \frac{z^C G(K,C)}{z^{C+1}} \right) dz = G(K,C) \cdot \frac{1}{2\pi i}\oint \frac{1}{z-0} dz = G(K,C) \tag{7.19}$$

wobei das Kreisintegral gegen den Uhrzeigersinn auf einem Kreis mit Radius $|z| < 1$ zu bilden ist. Das zweite Gleichheitszeichen in (7.19) ist Folge davon, daß bei der Integration alle Summanden der unendlichen Summe mit Ausnahme von $n = C$ verschwinden, das letzte Gleichheitszeichen ist Konsequenz der Anwendung von Cauchy's Integralsatz [Cop48]

$$f(\zeta) = \frac{1}{2\pi i}\oint \frac{f(z)}{z-\zeta} dz \tag{7.20}$$

auf die Funktion $f(z) \equiv 1$ an der Stelle $\zeta = 0$.

### 7.2.2 Grundidee der Approximation

Ausgangspunkt unserer Überlegungen war die Berechnung der stationären Wahrscheinlichkeit $\pi_{C-j}$ für die Anzahl $C-j$ belegter Kanäle auf der Ressource. Aus der Definitionsgleichung (7.12) geht hervor, daß sich diese Wahrscheinlichkeit mit Hilfe der Partitionsfunktionen $G(C)$ als

$$\pi_{C-j} = \frac{G(C-j) - G(C-j-1)}{G(C)} \tag{7.21}$$

darstellen läßt. Andererseits haben wir gesehen, daß sich die Partitionsfunktion gemäß (7.19) als komplexes Kreisintegral darstellen läßt. Die dort noch vorhandene Abhängigkeit von der Quellenzahl $K$ können wir beseitigen, indem wir zum sogenannten Poisson-Limes $K_s \to \infty$, $p_s \to 0$ mit $K_s \cdot p_s = \nu_s = \text{const}$ übergehen, der einem Poisson-Ankunftsprozeß entspricht. Hierfür gilt dann

$$\prod_{s=1}^{S}(1-p_s+p_s z^{d_s})^{K_s} \to \exp\left(\sum_{s=1}^{S}\nu_s(z^{d_s}-1)\right) \tag{7.22}$$

d. h. die Partitionsfunktion entspricht schließlich dem Kreisintegral

$$G(C) = \frac{1}{2\pi i}\oint_{|z|<1}\frac{\exp\left(\sum_{s=1}^{S}\nu_s(z^{d_s}-1)\right)}{z(1-z)^{C+1}}dz. \tag{7.23}$$

Eine Approximation der stationären Wahrscheinlichkeitsverteilung läßt sich demnach auf die Approximation dieses Kreisintegrals reduzieren. Ein Resultat von [Ble66] erlaubt letzteres durch Verwendung der sog. "*Sattelpunktmethode*". Hierzu definieren wir die Funktion

$$F(z) = \sum_{s=1}^{S}\nu_s(z^{d_s}-1) - C\log z, \tag{7.24}$$

die eng mit dem Integranden von (7.24) zusammenhängt; weiterhin sei der stationäre Punkt $z^*$ von $F$ die eindeutig bestimmte positive reelle Lösung der Gleichung $F'(z) = 0$ (für den Fall einer einzigen Klasse $s = 1$ beispielsweise entspricht $z^* = C/(\nu_1 d_1)$ dem Verhältnis zwischen Kapazität und mittlerer Anzahl belegter Kanäle).

Der Kern der Approximation besteht nun darin, daß die Verwendung der sogenannten Sattelpunktlinie $|z| = z^*$ für Berechnung des Kreisintegrals (7.23) es erlaubt, das Resultat der Berechnung als Funktion in $z^*$ anzugeben[1], und zwar gilt für eine "hinreichend vernünftige"[2] komplexe Funktion $h(z)$

$$\frac{1}{2\pi i}\oint_{|z|=z^*}h(z)\exp(F(z)) = \frac{\exp(F(z^*))}{\sqrt{2\pi F''(z^*)}}\left[h(z^*)+O\left(\frac{1}{C}\right)\right] \tag{7.25}$$

Für $h(z) = \dfrac{1}{z(z-1)}$ erhalten wir damit genau das gesuchte Integral (7.23). Für die Details der Berechnung verweisen wir auf [MM94].

Die letzten Endes resultierende Approximation (UAA) für die stationären Wahrscheinlichkeiten hat dann die Form

$$\pi_{C-j} = B\cdot(z^*)^j \tag{7.26}$$

---

1. Oberflächlich gesehen kann man dieses Vorgehen vielleicht als analog zur Taylorentwicklung auffassen, bei der ja auch eine Approximation damit erzielt wird, daß man die Funktion um einen günstig gewählten Punkt $z^*$ herum entwickelt.

2. Genauer gesagt muß sie analytisch ([BS87] S. 370ff.) auf einem Gebiet sein, das den Kreis $|z| = z^*$ enthält.

wobei $B = B(z^*)$ konstant in $j$ ist. Eine weitere Verfeinerung der Vorgehens im Verein mit einer noch deutlich länglicheren Rechnung[3] ermöglicht eine noch weitergehende Approximation (RUAA) der Form

$$\pi_{C-j} = (B_{(1)} + (j-1)B_{(2)} + (j-1)(j-2)B_{(3)}) \cdot (z^*)^j, \tag{7.27}$$

wiederum mit Termen $B_{(1)}, B_{(2)}, B_{(3)}$ die nur von $z^*$, aber nicht von $j$ abhängen. An dieser Stelle verzichten wir auf eine vertiefte Darstellung, geben dafür aber im folgenden Abschnitt 7.2.3 die für die numerischen Berechnungen benötigten Formeln vollständig an. Es bleibt noch anzumerken, daß diese Approximation Gültigkeit für $C \gg 1$, $\nu_s = O(C)$ und $j = O(1)$ hat. Ihr unschlagbarer Vorteil liegt in der niedrigen Komplexität, die in $O(1)$ bleibt, verglichen mit $O(C)$ für die bekannten Kaufman-Roberts-Rekursionen ([Kau81], [Rob82]) und noch deutlich höherer Komplexität für die direkte Berechnung, z.B. aus Abschnitt 7.1.

### 7.2.3 Algorithmische Darstellung

Wie angekündigt folgt nun die algorithmisch aufgebaute Darstellung sämtlicher Berechnungsschritte der beiden Approximationen.

#### Uniform Asymptotic Approximation (UAA)

Die Berechnung einer UAA benötigt folgende Schritte:

**Schritt 0:** Definiere die komplexe Funktion

$$F(z) = \sum_{s}^{S} \nu_s (z^{d_s} - 1) - C \log z \tag{7.28}$$

Die eindeutige Lösung $z^* > 0$ von

$$\frac{d}{dz} F(z) = 0 \tag{7.29}$$

kann mittels Newton-Verfahren oder Bisektion leicht berechnet werden, da sie

$$\sum_{s=1}^{S} \nu_s d_s (z^*)^{d_s} = C \tag{7.30}$$

erfüllt.

3. bei der selbst die zugehörigen Veröffentlichungen nicht ganz fehlerfrei sind (vgl. den Kasus [MRM98] Gleichung (4.30) versus [MMR99] Gleichung (4.26) – erst nochmaliges Nachrechnen ergab die Richtigkeit der letzteren Version!)

**Schritt 1:** Berechne

$$V = \sum_{s=1}^{S} \nu_s d_s^{\,2} (z^*)^{d_s}. \tag{7.31}$$

**Schritt 2:** Falls $z^* \neq 1$, dann ist $K$ (mit (7.31)) definiert als

$$K = \frac{1}{1-z^*} - \frac{\sqrt{V}\,\mathrm{sgn}(1-z^*)}{\sqrt{-2F(z^*)}}, \tag{7.32}$$

für $z^* = 1$ aber als

$$K = \frac{1}{2} + \frac{1}{6V} \sum_{s=1}^{S} \nu_s d_s^{\,3}. \tag{7.33}$$

Außerdem benötigen wir noch die "Complementary Error Function" $Erfc$ mit

$$Erfc(x) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-\xi^2} d\xi. \tag{7.34}$$

zur Berechnung von

$$M = \frac{1}{2} Erfc(\mathrm{sgn}(1-z^*)\sqrt{-F(z^*)}) + \frac{K e^{F(z^*)}}{\sqrt{2\pi V}} \tag{7.35}$$

**Schritt 3:** Berechne nun noch

$$B = \frac{e^{F(z^*)}}{M\sqrt{2\pi V}}. \tag{7.36}$$

Dann erhält man für die UAA der stationären Wahrscheinlichkeit, daß $C - j$ Kanäle einer Ressource mit einer Gesamtkapazität von $C$ Kanälen belegt sind:

$$\pi_{C-j}(C) = B(z^*)^j. \tag{7.37}$$

### Refined Uniform Asymptotic Approximation (RUAA)

RUAA verbessert UAA durch Approximation von $\pi_{C-j}(C)$ in höherer Ordnung, wobei einige der bereits angegebenen Formeln unverändert übernommen werden.

Der Algorithmus lautet damit folgendermaßen:

**Schritt 0:** wie oben.

**Schritt 1:** Berechne

$$T = \sum_{s=1}^{S} \nu_s {d_x}^2 (d_s - 3)(z^*)^{d_s} \tag{7.38}$$

$$Y = \sum_{s=1}^{S} \nu_s {d_x}^2 ({d_s}^2 - 6d_s + 11)(z^*)^{d_s} \tag{7.39}$$

$$W = \sum_{s=1}^{S} \nu_s {d_x}^2 ({d_s}^3 - 10{d_s}^2 + 35d_s - 50)(z^*)^{d_s} \tag{7.40}$$

**Schritt 2:** Falls $z^* \neq 1$, dann definiere

$$E = \frac{1 - 3z^* + 3(z^*)^2}{(1 - z^*)^3} + \frac{T(1 - 2z^*)}{2V(1 - z^*)^2} + \frac{1}{8}\left(\frac{5T^2}{3V^2} - \frac{Y}{V}\right)\frac{1}{1 - z^*} - \frac{V^{3/2}\,\mathrm{sgn}(1 - z^*)}{[-2F(z^*)]^{3/2}} \tag{7.41}$$

für $z^* = 1$ aber als

$$E = 1 + \frac{T}{2V} + \frac{5T^2}{24V^2} + \frac{35T^3}{432V^3} - \frac{Y}{8V}\left(1 + \frac{5T}{6V}\right) + \frac{W}{40V}. \tag{7.42}$$

**Schritt 3:** Unter Verwendung aller vorangegangenen Berechnungen definiere

$$B = \frac{e^{F(z^*)}}{M\sqrt{2\pi V}} \tag{7.43}$$

sowie

$$B^{(1)} = B + \frac{B}{8}\left(\frac{Y}{V^2} - \frac{5T^2}{3V^3}\right) + \frac{B^2 E}{V} \tag{7.44}$$

$$B^{(2)} = \frac{BT}{2V^2} \tag{7.45}$$

$$B^{(3)} = -\frac{B}{2V}. \tag{7.46}$$

Dann erhält man für die RUAA der stationären Wahrscheinlichkeit, daß $C - j$ Kanäle einer Ressource mit einer Gesamtkapazität von $C$ Kanälen belegt sind:

$$\pi_{C-j}(C) = [B^{(1)} + (j-1)B^{(2)} + (j-1)(j-2)B^{(3)}](z^*)^j. \tag{7.47}$$

### 7.2.4 Zusammenfassung

Der vorliegende Abschnitt diente der Einführung in die grundlegenden Elemente der Uniform Asymptotic Approximation UAA und ihrer Erweiterung RUAA zur Approximation der stationären Zustandswahrscheinlichkeiten eines Markov-Prozesses, der den Zustand einer einzelnen Verbindung der Kapazität $C$ im Fall von Mehrklassenverkehr beschreibt. UAA bzw. RUAA haben die in (7.37) bzw. (7.47) angegebene einfache Form und sind asymptotisch exakt für sehr große Kapazitäten, Ankunftsraten in der Größenordnung der Kapazität und Auslastungen nahe der Kapazitätsgrenze. Die numerische Anwendung dieser Approximationen auf unsere Fragestellung ist neu, daher wurden die hierfür benötigten Formeln übersichtlich in algorithmischer Form dargestellt. Als größter Vorteil dieser Approximationsmethode stellt sich ihre Komplexität heraus, die im Gegensatz zu bisher verwendeten Verfahren wie der Kaufman-Roberts-Rekursion nicht mehr von der Kapazität der Verbindung abhängt.

Abschließend sei bemerkt, daß sich die numerische Stabilität der Approximation weiter verbessern läßt, indem man beispielsweise die Berechnung von $K$ gemäß (7.32) bzw. (7.33) und von $E$ nach (7.41) bzw. (7.42) verfeinert. Hierzu wurde in [MRM98] eine nochmalige Erweiterung der Approximation hergeleitet, auf deren Einsatz im Rahmen der vorliegenden Untersuchungen aus Gründen der Übersichtlichkeit jedoch verzichtet wurde. Der Vollständigkeit halber sind die entsprechenden Formeln in Anhang A.4 wiedergegeben.

## 7.3 Validierung und Resultate

### 7.3.1 Der Referenzfall

UAA und RUAA wurden zur asymptotischen Approximation von stationären Zustandswahrscheinlichkeiten entwickelt; das hinsichtlich ihrer tatsächlichen numerischen Genauigkeit vorliegende Material ist noch sehr beschränkt (vgl. [MRM98]) und zudem auf das ursprüngliche Einsatzgebiet, nämlich die Planung von Virtual Private Networks, zugeschnitten und damit wenig aussagekräftig für unsere Fragestellung. Daher wurde in einem ersten Schritt eine Validierung der Approximation anhand der exakten Ergebnisse von Abbildung 7-4 unternommen. Das dort angegebene Szenario mit Kapazität $C = 120$ und Ankunftsraten $\nu_1 \in \{40, 80, 100, 120\}$ wird im folgenden der Einfachheit halber als Referenzfall bezeichnet.

Man beachte, daß in diesem Referenzfall zwar zwei unterschiedlich priorisierte Verkehrsklassen vorliegen, nämlich A-Gespräche und B-Gespräche, der zugrundeliegende Markovprozeß (Abbildung 7-2) die zweite Klasse jedoch trickreich umgeht, indem er deren Ankunftsprozeß durch die Annahme einer deterministischen Ankunftsrate $\infty$ approximiert. Daher reicht es aus, UAA/RUAA für den Fall $s = 1$, also einer priorisierten Klasse, anzuwenden[4]. Abbildung 7-5

zeigt den Zusammenhang zwischen den exakten Lösungen des Referenzfalles und den entsprechenden UAA/RUAA-Approximationen.

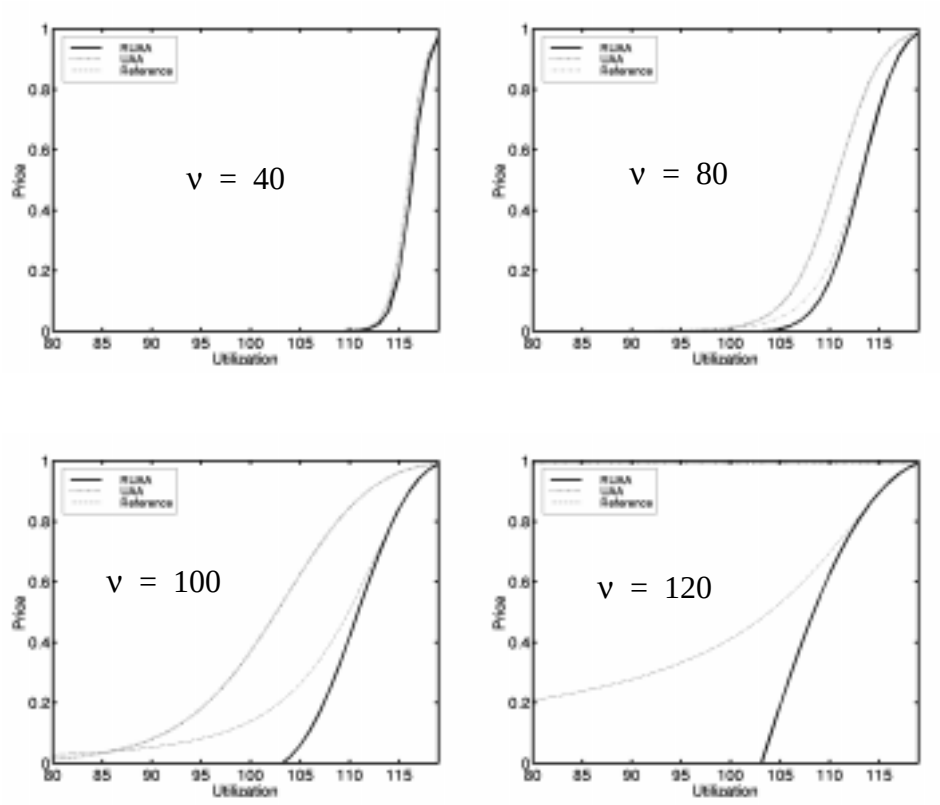


**Abbildung 7-5:** Preisfunktionen für $C = 120$: Exakte Lösung, UAA und RUAA für $\nu \in \{40, 80, 100, 120\}$

Hält man sich vor Augen, daß die Approximationen für den Fall sehr hoher Kapazitäten und Ankunftsraten sowie Auslastungen nahe der Kapazitätsgrenze entwickelt wurden, so stellt sich das Ergebnis als angenehme Überraschung heraus. Ist der ankommende Verkehr niedrig ($\nu = 40$, oben links), so fallen alle drei Kurven mehr oder weniger zusammen. Für mittleren Verkehr ($\nu = 80$ und $\nu = 100$, oben rechts bzw. unten links) liegt die exakte Lösung zwischen den beiden Approximationskurven, wobei RUAA unterhalb und UAA oberhalb der exakten Kurve bleiben. Dies bleibt auch für die gesättigte Verbindung der Fall ($\nu = 120$, unten rechts), auch wenn sich die Approximationen mehr und mehr von der exakten Lösung wegbewegen.

Allgemein läßt sich jedoch festhalten, daß zumindest die RUAA selbst im Falle der vergleichsweise kleinen Kapazität des Referenzfalles für Auslastungen nahe der Kapazität von der exak-

4. Hierbei wird bewußt die Existenz der “deterministischen” Verkehrsklasse unterschlagen. Zur Vereinfachung der Nomenklatur werden wir diesem Schema folgen und in Zukunft die Szenarien stets mittels der Anzahl “*stochastischer*” Klassen (also solcher mit stochastischem Ankunftsprozeß) charakterisieren und stillschweigend unterstellen, daß stets jeweils noch eine (“*deterministische*”) Klasse hinzukommt, die instantan beliebig viele Gespräche zur Verfügung stellen kann.

ten Lösung kaum zu unterscheiden ist. Dieser Bereich der guten Übereinstimmung verkleinert sich mit wachsendem Verkehrsaufkommen, parallel dazu vergrößert sich auch der Unterschied zwischen UAA und RUAA. Daraus läßt sich für die im folgenden untersuchten Szenarien mit großen Kapazitäten, für die die exakte Lösung nicht mehr berechenbar ist, immerhin die Schlußfolgerung ziehen, daß die RUAA umso besser mit der exakten Lösung übereinstimmt, je größer ihrerseits die Übereinstimmung zwischen UAA und RUAA ist.

### 7.3.2 Große Kapazitäten

In einem nächsten Schritt wurden die beiden Approximationen für den Fall von Verbindungskapazitäten berechnet, wie sie in heutigen Hochgeschwindigkeitsnetzen von Bedeutung sind.

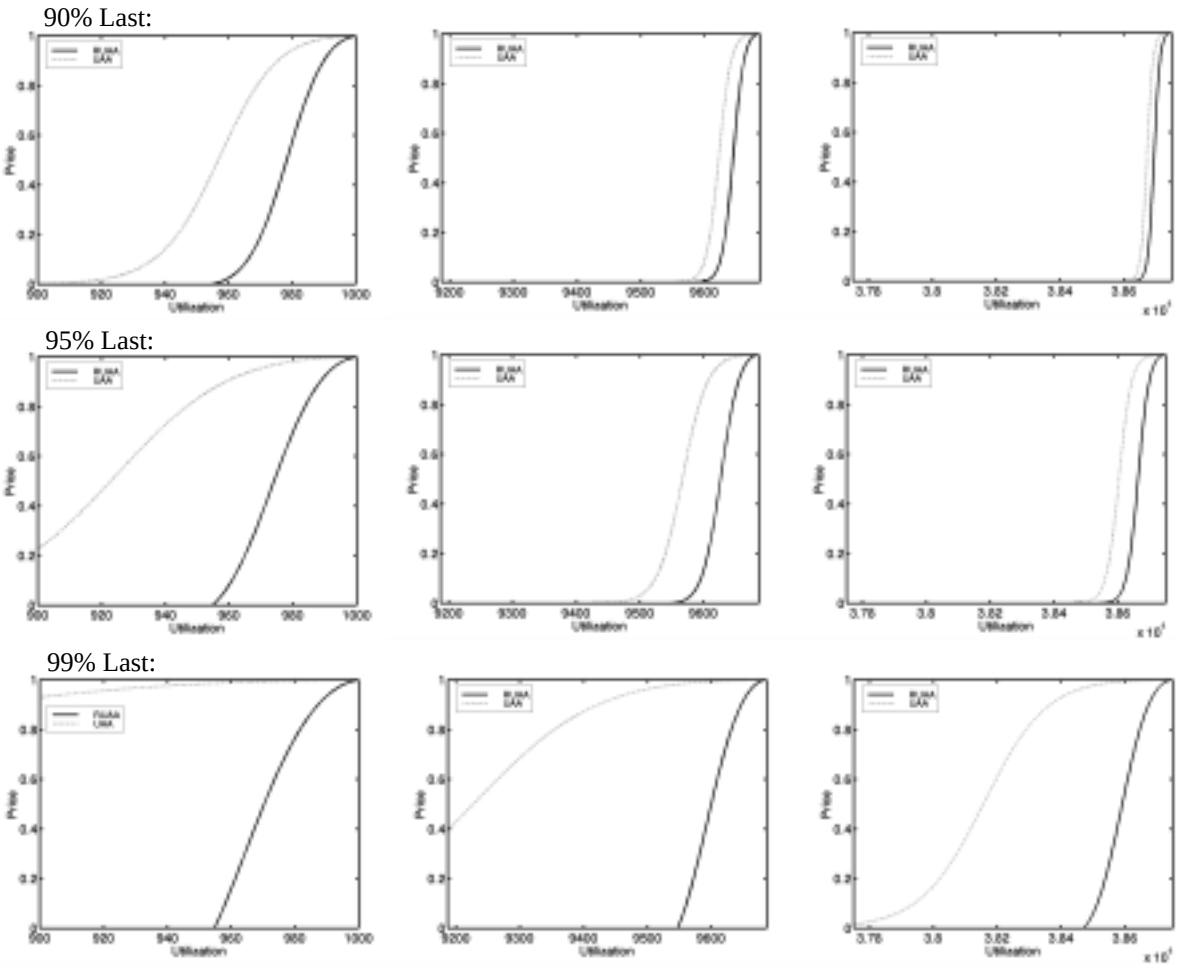


**Abbildung 7-6:** Preisfunktionen für große Kapazitäten ($C \in \{1000, 9687, 38750\}$), eine stochastische Klasse und 90%, 95% bzw. 99% Last

Nimmt man einmal an, daß eine IP-Telefonverbindung niedriger Sprachqualität (wie sie in etwa einem herkömmlichen Telefongespräch entspricht) eine Übertragungsrate von 16 kbps erfordert (was im folgenden einem “Kanal” als kleinster Kapazitätseinheit gleichgesetzt wird),

dann lassen sich auf einer 155 Mbps-Verbindung etwa 9687 Kanäle, auf einer 622 Mbps-Verbindung sogar ca. 38750 Kanäle unterbringen. Dennoch bleiben die entsprechenden Approximationen berechenbar. Abbildung 7-6 zeigt die Ergebnisse von UAA und RUAA für Kapazitäten $C \in \{1000, 9687, 38750\}$ und 90%, 95% bzw. 99% Last (wobei 90% Last einer Poissonrate für Ankünfte der stochastischen Verkehrsklasse in Höhe von 90% der Kapazität entspricht usw.).

Abbildung 7-6 zeigt einmal, daß für gegebene Last die Kurven von UAA und RUAA mit steigender Kapazität einander näherrücken, was nach den Ausführungen am Ende von Abschnitt 7.3.1 darauf schließen läßt, daß die Güte der Approximation durch RUAA steigt und etwa im Fall des oben rechts abgebildeten Szenarios ($C = 38750$, 90% Last) RUAA und exakte Kurve mehr oder weniger zusammenfallen dürften (analog zu Abbildung 7-5 obere Reihe). Zweitens wird die Approximation mit steigender Verkehrslast ungenauer, was der Gewinn an Approximationsgüte durch steigende Kapazitäten allerdings wieder ausgleichen kann (z.B. läßt die Ähnlichkeit der Kurven für $C = 1000$/95% Last (Mitte links) und $C = 38750$/99% Last (unten rechts) auf eine in etwa vergleichbare Approximationsgüte schließen).

### 7.3.3 Mischklassen-Szenarien

Neben der Ausdehnung unseres Modells auf große Kapazitäten ermöglicht die Verwendung von UAA/RUAA aber auch noch die Behandlung von Verkehr, der aus einer Menge von Klassen mit unterschiedlichem Ankunftsverhalten, Verweilzeiten und Bandbreitenanforderungen besteht. In diesem Abschnitt betrachten wir hierzu die Approximation von Modellen, die eine Preisfunktion für die deterministische Verkehrsklasse berechnen, wenn die stochastische Klasse in Wirklichkeit aus unterschiedlich charakterisierten Unterklassen zusammengesetzt ist. Anders ausgedrückt besteht die stochastische Klasse aus einer Mischung von ankommenden Gesprächen, die zwar unterschiedliche Bandbreitenanforderungen, Ankunftsraten und Verweilzeiten aufweisen können, aber dennoch in identischer Weise priorisiert zu behandeln sind und auch alle jeweils den Einheitspreis für stochastische Gespräche, nämlich 1, entrichten (in diesem Fall natürlich noch skaliert durch die Bandbreite).

Der Vorteil dieses Mischklassen-Konstruktes besteht darin, daß die Trunk Reservation Policy immer noch optimal hinsichtlich der Einnahmenmaximierung bleibt. In Abschnitt 7.1.4 haben wir bereits gesehen, daß das Hauptproblem hierbei vielmehr im Übergang von einem “truncated birth-and-death-process” zu einem allgemeineren “truncated Markov process” liegt. Letztere lassen sich jedoch mit UAA/RUAA problemlos behandeln. Lediglich (7.5) bedarf einer kleineren Anpassung: für Trunk Reservation-Parameter $a$ kommt im Mischklassenszenario aufgrund der bekannten Additivität von Poissonprozessen [Hav98] neuer Verkehr mit Poissonrate $\sum \nu_s$ an und benötigt durchschnittlich $(\sum d_s \nu_s)/(\sum \nu_s)$ Kanäle, was insgesamt auf einen Fluß von

$$x_a = \left( \sum_{s=1}^{S} d_s \nu_s \right) (1 - \pi_a(C)) \tag{7.48}$$

führt.

Demgegenüber bedarf es keiner Änderung von (7.4), da die unterschiedlichen Verweilzeiten über den Begriff der "Verkehrsintensität" aufgefangen werden (siehe Abschnitt 7.2.1), daher bleibt die Rate, mit welcher der Zustand $C - a$ nach unten hin verlassen wird, gleich $(C - a) \cdot \pi_a(C - a)$.

Allerdings kann die Zulassung bereits einer noch ungemischten stochastischen Verkehrsklasse, die mehr als einen Kanal Bandbreite anfordert, zu Seiteneffekten führen, wie sie in Abbildung 7-7 für UAA (links) und RUAA (rechts) dargestellt sind. Man kann beobachten, daß im Sättigungsfall die Preisfunktion nur für die Anforderung von einem Kanal auf 1 ansteigt. Sobald die stochastische Klasse zwei oder mehr Kanäle benötigt, kann es vorkommen, daß die nahezu gesättigte Verbindung nurmehr weniger freie Kanäle aufweist als ein einziges Gespräch der stochastischen Klasse benötigt. In diesem Fall wäre es für die Auslastung günstiger, die freie Kapazität mit Gesprächen der deterministischen Klasse zu füllen (die ja immer nur einen Kanal pro Gespräch benötigen); demzufolge ist der Preis für ein derartiges Gespräch echt kleiner als 1 und sinkt darüberhinaus mit wachsender Größe der Bandbreitenanforderung.

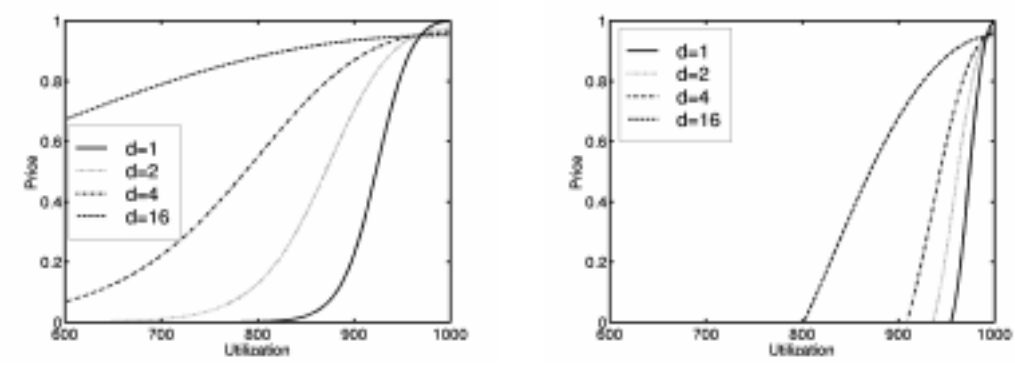


**Abbildung 7-7:** Seiteneffekte in Szenarien mit einer stochastischen Klasse, die jedoch mehr als einen Kanal Bandbreite anfordern kann: UAA (links) und RUAA (rechts) für eine Verbindung mit 1000 Kanälen, 95% Last und einer stochastischen Klasse, die 1, 2, 4 oder 16 Kanäle pro Gespräch anfordert.

Abbildung 7-8 zeigt die Preisfunktion für die deterministische Verkehrsklasse im Fall einer 155 Mbps-Verbindung, wobei die (stochastische) Mischklasse zusammengesetzt ist aus Sprachverkehr niedriger (16 kbps) oder hoher (64 kbps) Qualität sowie Videoverkehr (384 kbps). Szenario A (links) geht von 95% Last aus und setzt die Mischklasse zur Hälfte aus

schlechtem Sprachverkehr und zur anderen Hälfte aus gutem Sprachverkehr zusammen. Szenario B (rechts) weist bei wiederum 95% Last die Hälfte des Mischklassenverkehrs Videoverbindungen zu und teilt die andere Hälfte gleichermaßen zwischen gutem und schlechtem Sprachverkehr auf.

.

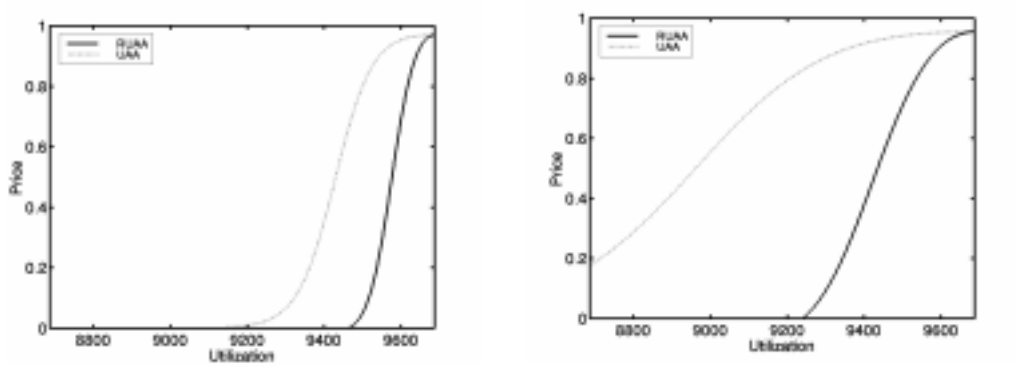


**Abbildung 7-8:** Preisfunktionen für Mischklassen-Szenarien mit Mischklasse bestehend aus zwei Klassen (links) bzw. drei Klassen (rechts), jeweils UAA und RUAA.

## 7.4 Preisfunktionen im Mehrklassenfall

Nachdem wir in Abschnitt 7.3.3 das Instrumentarium bereitgestellt haben, um die Preisfunktion für die deterministische Klasse zu berechnen, wenn die stochastische Klasse aus einer Mischung verschieden charakterisierter Unterklassen besteht, untersuchen wir nun einen Ansatz, um für Verkehrsszenarien mit *mehreren stochastischen Klassen* Preisfunktionen für jede einzelne dieser stochastischen Klassen zu berechnen. Hierzu wird zunächst dargestellt, wie sich durch geschickte Zusammenfassung einzelner stochastischer Klassen verschiedene Mischklassen-Szenarien ergeben, die mit den Mitteln des vorhergehenden Abschnittes gelöst werden können. Im Anschluß daran wird exemplarisch die Lösung des dadurch entstehenden linearen Gleichungssystems demonstriert.

### 7.4.1 Mehrklassen-Szenario als Folge von Mischklassen-Szenarien

Nehmen wir also an, der vorliegende Verkehr setze sich aus $S$ stochastischen Klassen zusammen, die jeweils verschiedene Bandbreitenanforderungen und Verkehrsintensitäten aufweisen können. Weiterhin bezahle ein Gespräch der höchsten Klasse 1 nach wie vor einen Betrag von $p_1 = 1$ (wiederum pro Kanal), wohingegen Verkehr der Klassen 2, ..., $S$ entsprechend jeweils $p_2, \ldots, p_S$ mit $1 \geq p_2 \geq \ldots \geq p_S$ bezahle. Darüberhinaus unterscheide sich Klasse Nummer $S+1$ von den anderen durch ihren “deterministischen” Ankunftsprozeß mit Rate $\infty$, ihrer

Bandbreitenanforderung von nur einem Kanal und ihrem Preis $p_{S+1}$, der niedriger als alle anderen ist.

Wie in Abschnitt 7.1.4 bereits angerissen, besteht das Hauptproblem bei der Bestimmung der korrekten Werte für $p_2, \ldots, p_S$ darin, daß im Mehrklassenfall nicht länger davon ausgegangen werden kann, daß Trunk Reservation die Maximierung der Einnahmen aus der Verbindung gewährleistet. Um dieses Problem zu umgehen, schlagen wir einen Ansatz vor, der zumindest die näherungsweise Bestimmung der gesuchten Preise ermöglicht. Die Grundidee nützt hierbei nachhaltig die Tatsache aus, daß sich aus der Kombination zweier Poissonprozesse jeweils wieder ein neuer Poissonprozeß mit entsprechend höherer Rate ergibt. Andererseits resultiert aus der Kombination eines Poissonprozesses mit dem beschriebenen deterministischen Ankunftsprozeß der Rate $\infty$ unverändert ein deterministischer Prozeß mit unendlicher Rate. Damit läßt sich durch geeignete Kombination von Verkehrsklassen schließlich immer ein lösbares Mischklassenszenario konstruieren. Aus der wiederholten Durchführung dieser Idee für verschiedene geeignete Klassenkombinationen (in Anlehnung an [Rei99a] und [Rei99b] auch *Pseudoklassen* genannt) läßt sich dann ein lineares Gleichungssystem bilden, dessen Lösung schließlich die gesuchten Preise für die einzelnen stochastischen Klassen liefert.

Betrachten wir also die oberen *S* Klassen mit ihren Poisson-Ankunftsraten von $\nu_s$. Faßt man diese Klassen künstlich zu einer Pseudoklasse $\langle S \rangle$ zusammen, so finden Ankünfte dieser Pseudoklasse immer noch als Poissonprozeß statt, nun allerdings mit Rate

$$\nu_{\langle S \rangle} = \sum_{s=1}^{S} \nu_s . \tag{7.49}$$

Weiterhin bezahlt diese Pseudoklasse einen Preis

$$p_{\langle S \rangle} = \left( \sum_{i=1}^{S} \nu_i p_i \right) \Big/ \left( \sum_{i=1}^{S} \nu_i \right) \tag{7.50}$$

d.h. ein gewichtetes Mittel der Preise für die individuellen stochastischen Klassen.

Andererseits erfüllt die Pseudoklasse $\langle S \rangle$ in Verbindung mit der gewöhnlichen deterministischen Klasse die Voraussetzungen an ein Mischklassenszenario gemäß Abschnitt 7.3.3. Deshalb lassen sich die dort entwickelten Methoden dazu verwenden, um einen Preis $\tilde{p}_{S+1} = p_{S+1} / p_{\langle S \rangle}$ für die Zulassung eines deterministischen Gesprächs zu berechnen (das nunmehr allerdings mittels $p_{\langle S \rangle}$ anstelle von 1 zu skalieren ist).

Analog dazu werden der Reihe nach für jedes $k = 1, \ldots, S-1$ die Klassen 1, 2, ..., *S*-*k* zur Pseudoklasse $\langle S-k \rangle$ zusammengefaßt, wobei jeweils

$$\nu_{\langle S-k \rangle} = \sum_{s=1}^{S-k} \nu_s \tag{7.51}$$

und

$$p_{\langle S-k \rangle} = \left( \sum_{i=1}^{S-k} \nu_i p_i \right) \Big/ \left( \sum_{i=1}^{S-k} \nu_i \right) \tag{7.52}$$

gilt, während die Klassen $S-k+1, \ldots, S+1$ zu einer deterministischen Pseudoklasse mit unendlich hoher Ankunftsrate zusammengefaßt werden. Das entstehende Mischklassenszenario erlaubt jeweils die Berechnung eines Preises für die Zulassung eines Gesprächs aus der deterministischen Klasse, bezogen auf $p_{\langle S-k \rangle}$. Zuletzt erhalten wir ein einfaches lineares Gleichungssystem zur Berechnung der Preise $p_2, \ldots, p_S$ für die einzelnen stochastischen Klassen.

### 7.4.2 Ein Beispiel

Betrachten wir der Einfachheit halber ein Szenario mit $S = 2$ stochastischen Klassen [Rei99a]. Verkehr der Klasse 1 kommt an als Poissonstrom mit Rate $\nu_1$ und benötigt Bandbreite $B_1$, Verkehr der Klasse 2 kommt an als Poissonstrom mit Rate $\nu_2$ und braucht Bandbreite $B_2$. Verkehr der Klasse 3 kommt dagegen "deterministisch" an mit Rate $\infty$ und benötigt Bandbreite 1. Verweilzeiten der Gespräche werden als exponentialverteilt mit Mittel 1 angenommen (wiederum als einfache Konsequenz der bereits angesprochenen "insensitivity property" von UAA/RUAA, die es erlaubt, unterschiedliche Verweilzeiten unter dem Begriff der Verkehrsintensität zu subsumieren). Darüberhinaus zahle ein Gespräch der Klasse 1 pro Bandbreiteneinheit (Kanal) einen Betrag von 1, ein Gespräch der Klasse 2 oder 3 entsprechend $p_2$ bzw. $p_3$.

Nehmen wir außerdem an, es gebe ein optimales Zulassungsschema $S_{opt}$, um die Einnahmen aus der Verbindung zu maximieren. Aufgrund dieses Schemas wird die Verbindung momentan von einem Tripel $(x_1, x_2, x_3)$ belegt, wobei $x_i$ für $i = 1, 2, 3$ jeweils die gesamte derzeit von Gesprächen der Klasse $i$ belegte Bandbreite bezeichnet. Daraus ergibt sich in Verbindung mit den eingeführten Preisen eine Gesamteinnahme von

$$R_{opt} = 1 \cdot x_1 + p_2 \cdot x_2 + p_3 \cdot x_3 \tag{7.53}$$

Wechselt man die Perspektive analog zu Abschnitt 7.1.3, so stellt sich jetzt die Frage, wie hoch die Preise $p_2$ bzw. $p_3$ sein müssen, damit es für den Betreiber der Verbindung keine Rolle spielt, ob er ein ankommendes Klasse 2- bzw. Klasse 3-Gespräch annimmt oder stattdessen auf die Ankunft eines (teureren) Klasse 1-Gesprächs wartet. Hierzu gehen wir in zwei Schritten vor:

**Schritt 1:** Wir fassen Klasse 1 und Klasse 2 zur Pseudoklasse $\widehat{12}$ mit Poisson-Ankunftsrate $\nu_1 + \nu_2$ zusammen. Diese Pseudoklasse bezahle im Mittel einen Betrag von $p_{\widehat{12}}$. Zusammen mit der deterministischen Klasse 3 liegt damit ein Mischklassenszenario vor, und wir können den Preis $p_3^*$ berechnen, den ein Klasse 3-Gespräch zahlen muß, wenn es angenommen werden möchte, obwohl gerade $x_1 + x_2$ Kanäle von Pseudoklasse $\widehat{12}$ belegt werden, und zwar jeweils zum Preis 1.

Andererseits beträgt der mittlere von Gesprächen der Pseudoklasse $\widehat{12}$ entrichtete Preis in Wirklichkeit

$$p_{\widehat{12}} = \frac{\nu_1}{\nu_1 + \nu_2} + p_2 \cdot \frac{\nu_2}{\nu_1 + \nu_2} \tag{7.54}$$

anstelle von 1. Daher ist $p_3^*$ noch mit diesem Faktor zu skalieren, und die resultierenden Gesamteinnahmen der Verbindung ergeben sich zu

$$R_{\widehat{12}} = p_{\widehat{12}} \cdot (x_1 + x_2) + p_3^* \cdot p_{\widehat{12}} \cdot x_3 \tag{7.55}$$

**Schritt 2:** Jetzt fassen wir Klasse 2 und Klasse 3 zur Pseudoklasse $\widehat{23}$ zusammen, deren Verkehr deterministisch mit Rate $\infty$ ankommt, während Klasse 1 immer noch ein Poissonstrom mit Rate $\nu_1$ ist. In diesem Szenario beträgt der mittlere Preis für die Akzeptanz eines Pseudoklassengesprächs jetzt $p_{\widehat{23}}$, und die resultierenden Gesamteinnahmen sind

$$R_{\widehat{23}} = x_1 + p_{\widehat{23}} \cdot (x_2 + x_3) \tag{7.56}$$

Bis jetzt haben wir also das originale Dreiklassenproblem reduziert auf zwei Mischklassenszenarien zur Berechnung der Preise $p_3^*$ and $p_{\widehat{23}}$. Andererseits führt jedes dieser drei Szenarien zur selben Gesamteinnahme

$$R_{\widehat{12}} = R_{\widehat{23}} = R_{opt}, \tag{7.57}$$

da man andernfalls für den originalen Zweiklassenfall ein Zulassungsschema angeben könnte, das besser als die Trunk Reservation Policy wäre, indem man den Fluß an A-Gesprächen künstlich in zwei unterschiedliche Klassen aufteilt und diese dann gemäß $S_{opt}$ behandelt (im Widerspruch zur bekannten Optimalität der Trunk Reservation Policy). Folglich ergibt sich aus (7.53), (7.55), (7.56) und (7.57) das folgende lineare Gleichungssystem:

$$\begin{aligned} R_{opt} &= x_1 + p_2 \cdot x_2 + p_3 \cdot x_3 \\ &= p_{\widehat{12}} \cdot (x_1 + x_2) + p_3^* \cdot p_{\widehat{12}} \cdot x_3 \\ &= x_1 + p_{\widehat{23}} \cdot (x_2 + x_3) \end{aligned} \tag{7.58}$$

Hier sind $p_2$ und $p_3$ die Unbekannten, $p_3^*$ und $p_{\widehat{23}}$ werden gemäß Schritt 1 bzw. Schritt 2 berechnet, und der Wert von $p_{\widehat{12}}$ hängt nur von $p_2$ ab, wie in (7.54) beschrieben. Die Lösung von (7.58) ergibt schließlich die korrekten Werte für $p_2$ und $p_3$.

Abbildung 7-9 zeigt exemplarisch das Aussehen der resultierenden Lösung für eine Verbindung der Kapazität 1000 Kanäle mit hoch- und minderwertigem Sprachverkehr, der insgesamt 90% der Kapazität belegt, und zwar jeweils zur Hälfte.

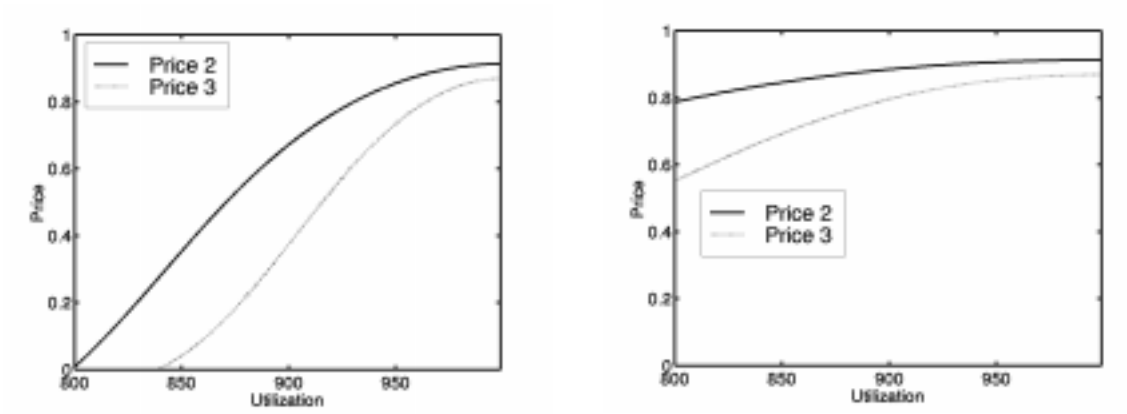


**Abbildung 7-9:** Preise $p_2$ und $p_3$ für das Beispiel mit zwei stochastischen Klassen: $C = 1000$, $(x_1, x_2) = (450, 450) = (\nu_1 B_1, \nu_2 B_2)$, $(B_1, B_2, B_3) = (16, 4, 1)$. RUAA (links) und UAA (rechts).

### 7.4.3 Zur Abhängigkeit von der momentanen Kapazitätsaufteilung

Eine Eigenheit des vorgestellten Ansatzes besteht darin, daß die sich ergebenden Preise einmal von den durchschnittlichen Ankunftsraten $\nu_s$ abhängen, andererseits aber auch von der aktuell gerade gültigen Aufteilung $(x_1, \ldots, x_{S-1})$ des Verkehrsaufkommens, die ja aufgrund statistischer Schwankungen nicht von vorneherein mit den durchschnittlichen Ankunftsraten übereinstimmt. [Sp99] hat den Einfluß dieser aktuell gemessenen Aufteilung auf die Preisfunktionen untersucht und kommt zu folgenden Ergebnissen: Ist die Auslastung der Verbindung hoch, dann hat eine Änderung der Poisson-Ankunftsraten bei fixierter gemessener (realer) Kapazitätsaufteilung kaum Einfluß auf die Preisfunktionen. Dasselbe gilt auch für weniger hohe Auslastungen im Fall von $p_2$, während $p_3$, der Preis für die deterministische Klasse, durchaus Variationen zeigt. In jedem Fall ist $p_2$ höher als $p_3$.

Abbildung 7-10 zeigt ein typisches Beispiel für die geschilderten Verläufe. Es handelt sich um ein Szenario mit zwei stochastischen Klassen (analog Abbildung 7-9, $C = 1000$, $(B_1, B_2, B_3) = (16, 4, 1)$). Von den 1000 Kanälen sind jeweils insgesamt 900 von den beiden stochastischen Klassen belegt. Die erste Spalte zeigt die Preisfunktionen, wenn 10% dieser 900 Kanäle von Klasse 1 belegt wird und 90% von Klasse 2, die zweite Spalte entspricht einer rea-

len Aufteilung der 900 Kanäle im Verhältnis 25 zu 75, die dritte Spalte schließlich einem Verhältnis von 50 zu 50. Orthogonal dazu wird in den Reihen die jeweilige Ankunftsrate variiert: Reihe 1 entspricht $(\nu_1 B_1, \nu_2 B_2) = (90, 810)$, Reihe 2 $(\nu_1 B_1, \nu_2 B_2) = (225, 675)$, Reihe 3 $(\nu_1 B_1, \nu_2 B_2) = (450, 450)$, Reihe 4 $(\nu_1 B_1, \nu_2 B_2) = (675, 225)$ und die unterste Reihe schließlich $(\nu_1 B_1, \nu_2 B_2) = (810, 90)$. Für die Auswertung weiterer Parameterkombinationen sei nochmals auf [Sp99] verwiesen.

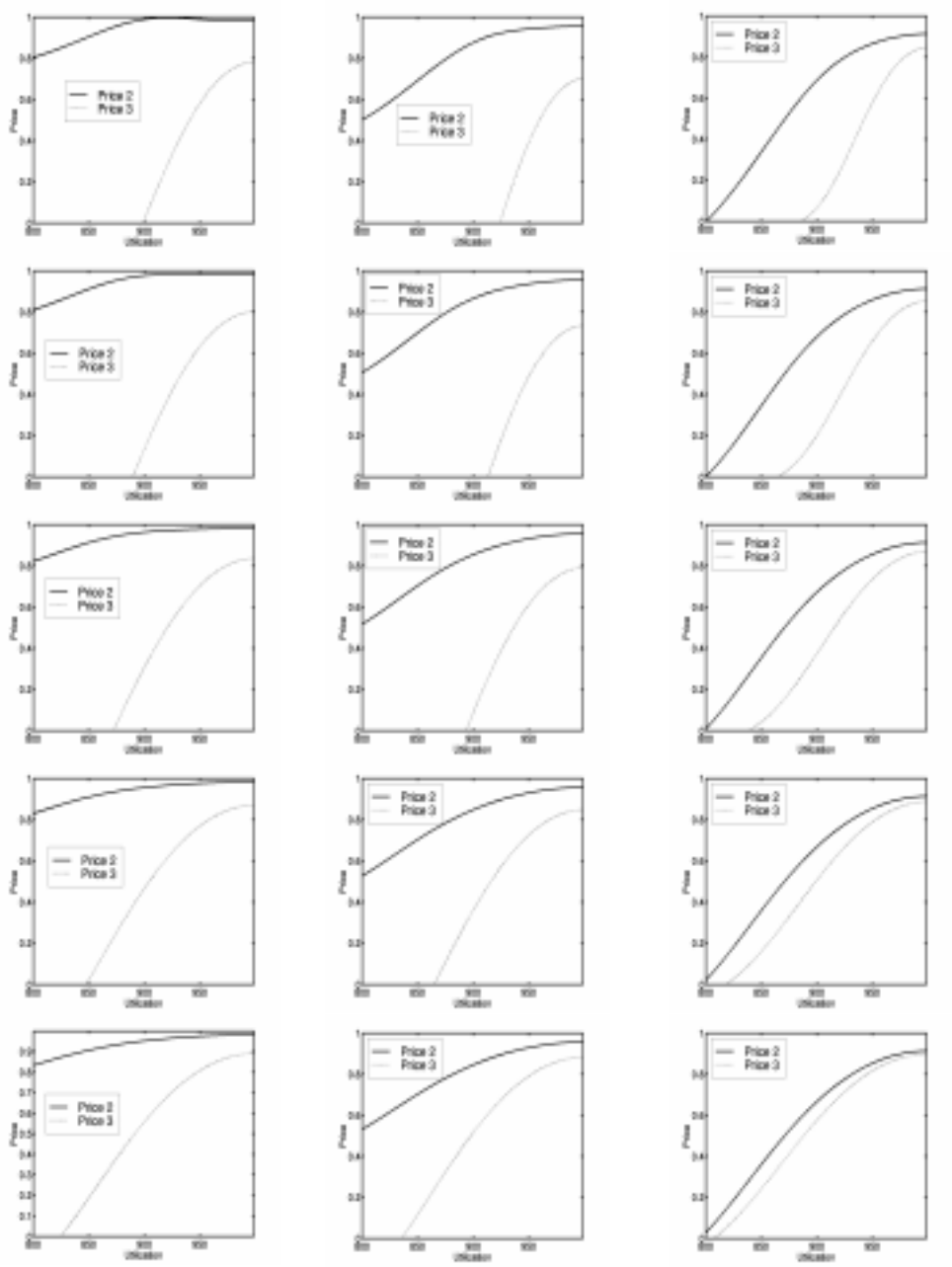


**Abbildung 7-10:** Abhängigkeit der Preisfunktionen von Ankunftsraten und real vorliegender Kapazitätsaufteilung: RUAA für zwei stochastische Klassen. Horizontal variiert der reale (gemessene) Verkehr, vertikal die Ankunftsraten der entsprechenden Poissonprozesse.

## 7.5 Zusammenfassung, Interpretation und Ausblick

Zum Abschluß dieses Kapitels fassen wir noch einmal kurz die einzelnen Schritte des vorgestellten Modellierungsansatzes zusammen und zeigen gleichzeitig Richtungen auf, in die sie jeweils im Hinblick auf die Verwendung in einem Charging- and Accounting-Tool zu interpretieren sind.

Allgemein ist zunächst festzuhalten, daß das eingeführte und erweiterte Modell ursprünglich aus der Welt der Telefonie kommt. Da jedoch insbesondere das IntServ-Framework, wie es im Detail bereits in Abschnitt 6.3.1 erläutert wurde, auf einem Flow-Konzept basiert, das (unter der Annahme, daß jeder Verkehr jeweils ein einzelner Flow ist) den in der Telefonie verwendeten Konzepten von Gesprächen, Kanälen etc. mehr oder weniger äquivalent ist, lassen sich auch die entsprechenden mathematischen Modelle aus dem Gebiet der Stochastischen Netzwerke auf entsprechende Situationen im Internetverkehr übertragen.

### Kelly's Bound

Ausgangspunkt war Kelly's Bound im Spezialfall einer betrachteten Verbindung geringer Kapazität, auf der zwei unterschiedlich priorisierte, aber hinsichtlich ihrer QoS-Anforderungen ununterscheidbare Verkehrsklassen gleichzeitig bedient wurden. Unter der Annahme, daß A-Gespräche immer eine Gebühr von 1 entrichten, ließ sich in Abhängigkeit von der Auslastung der Verbindung die Gebühr ermitteln, die ein im Moment zur Verfügung stehendes B-Gespräch zahlen müßte, um vom Netzbetreiber anstelle eines irgendwann später ankommenden, aber lukrativeren A-Gesprächs akzeptiert zu werden. Der Unterschied zwischen den Gebühren für A- und B-Gespräche kompensiert also im wesentlichen die Tatsache, daß der Netzbetreiber durch das Warten auf ein A-Gespräch für eine gewisse Zeit den zur Verfügung stehenden Kanal leer lassen müßte. Das B-Gespräch könnte den Kanal dagegen sofort belegen und kann deswegen billiger sein als das A-Gespräch.

### Erste Erweiterung: Hohe Kapazitäten

Zunächst wurde durch Verwendung der Unified Asymptotic Approximation UAA bzw. der Refined Uniform Asymptotic Approximation RUAA die näherungsweise Lösung des beschriebenen Szenarios für im heutigen Internet realistische Kapazitäten ermöglicht, wobei die Anzahl der Verkehrsklassen weiterhin auf zwei beschränkt blieb. Das hiermit erreichte Resultat läßt zweierlei Interpretationsansätze zu:

- **Interpretation A:**

  Zum einen ist es durchaus denkbar (und verschiedene neuere Ergebnisse, z.B. aus dem Bereich der Portfolio-Theorien, legen dies auch nahe), daß die Anzahl unterschiedlicher Klassen im zukünftigen Internet auf zwei beschränkt bleibt. In diesem Fall stünde also eine "Premium"-Klasse mit hoher Priorität der bekannten "Best-Effort"-Klasse mit

ihrer ständigen Verfügbarkeit gegenüber. Setzt man nun noch eine Tarifstruktur voraus, in welcher der Premium-Verkehr unabhängig von der Netzauslastung eine volumenabhängige Tarifierung erfährt, dann läßt sich der durch unser Modell ermittelte Preis unmittelbar als der Tarif für den Best-Effort-Verkehr auffassen, mit dem insgesamt die Einnahmen des Netzbetreibers maximiert werden.

- **Interpretation B:**

  Der zweite denkbare Interpretationsansatz ist weniger unmittelbar, aber universaler verwendbar. Nehmen wir hierzu an, der untersuchte Verkehr sei von seinen Charakteristika her nicht weiter unterschiedlich klassifiziert. Für gegebene Netzauslastung ist ein Tarif für eine neu ankommende Verbindung gesucht, und zwar bezogen auf eine Bandbreiteneinheit (z.B. einen Kanal).

  Wir haben oben gesehen, daß in Kelly's Bound der Unterschied der Tarife darauf zurückzuführen ist, daß die Annahme von B-Gesprächen eine sofortige Auslastung der Leitung ermöglicht, während das Warten auf ein A-Gespräch eine Zeitlang mit gewisser Ressourcenverschwendung verknüpft ist, was die höheren Gebühren für A-Gespräche rechtfertigt. Dies läßt für unseren hier betrachteten Fall folgende Interpretation zu:

  Wir wollen den Preis für eine gerade vorliegende Verbindungsanfrage bestimmen. Dann können wir die Situation zunächst einmal dadurch "aufblasen", daß wir diese Verbindungsanfrage einer fiktiven B-Klasse (von immer verfügbaren Anfragen) zuordnen, während aller übriger Verkehr in einer entsprechenden fiktiven A-Klasse zusammengefaßt ist. In diesem Fall läßt sich dann der Tarif ermitteln, der die sofortige Verfügbarkeit der betrachteten Verbindung gegenüber allen später ankommenden Anfragen honoriert und in diesem Sinne einen brauchbaren auslastungsabhängigen Preis für die Verbindung angibt. Nach Ermittlung dieses Tarifes "kollabieren" die fiktiven Klassen wieder, um bei Ankunft der nächsten Verbindungsanfrage entsprechend neu initialisiert zu werden.

### Zweite Erweiterung: Mischklassen-Szenarien

Während wir bislang kurzerhand davon ausgingen, die vorhandene Auslastung als zusammengesetzt aus der Belegung von zunächst unabhängigen Einzelkanälen modellieren zu können, erlaubt es die Allgemeinheit von UAA und RUAA, tatsächlich die unterschiedlichen Bandbreitenanforderungen und Verbindungsdauern direkt zu berücksichtigen. Damit erlaubt ein derartiges Mischklassenszenario den weitergehenden Einsatz von Interpretation B, indem aller Verkehr außer der gerade angekommenen Verbindungsanfrage mit seiner von Gespräch zu Gespräch unterschiedlichen Charakteristik in einer Mischklasse zusammengeführt wird, die der fiktiven A-Klasse von eben entspricht, während die betrachtete Verbindungsanfrage nach dem "Aufblasen" der Situation als B-Gespräch interpretiert wird. Entsprechend kann nun wiederum der Tarif für die angekommene Anfrage unter Berücksichtigung der momentanen Netzauslastung bestimmt werden. Daß, wie in Abschnitt 7.3.3 ausgeführt, hierbei immer noch die

Optimalitätseigenschaften der Trunk Reservation zum Tragen kommen, begründet die Güte des hieraus resultierenden Preismodells.

### Dritte Erweiterung: Mehrklassen-Szenarien

In dieser letzten Erweiterung wurde ein Ansatz vorgestellt, der eine direkte Tarifierung mehrerer unterschiedlicher Verkehrsklassen im Sinne von Interpretation A zuließe. Hierzu wurde versucht, das Mehrklassen-Szenario näherungsweise als Sequenz von Zweiklassenszenarien darzustellen, deren jeweilige Lösung mit den bisherigen Mitteln möglich ist. Allerdings stellt sich heraus, daß die Ergebnisse in der vorliegenden Form (noch) nicht unmittelbar zur Verkehrstarifierung herangezogen werden können. Dies liegt unter anderem daran, daß es in diesem Fall darauf ankommt, was unter der unterschiedlichen Priorisierung verschiedener Verkehrsklassen genau zu verstehen ist.

### Ausblick

Die Antwort auf letzteres Problem ist nur in engem Kontakt mit der Praxis zu klären, wie dies im Rahmen des CATI-Projektes vorgesehen ist. In diesem Zusammenhang ist es allerdings komplett unrealistisch, bei jedem ankommenden Verbindungswunsch die in diesem Kapitel entwickelte komplexe Preisberechnungsmethode in Echtzeit durchzuführen. Vielmehr werden die vorgestellten Konzepte in Form von vorab berechneten “typischen” Preiskurven zum Einsatz kommen, die in den entsprechenden Knoten tabellarisch abgespeichert werden. Zur weiteren Reduktion der Komplexität wird zudem noch eine geeignete stückweise Linearisierung der Kurven vorgenommen. Durch die Vielzahl der so zusammenkommenden Approximationen wird der Kern des Modells allerdings nicht zerstört: Das resultierende Tarifschema gibt für eine betrachtete Verbindung lokal einen Preis an, der insbesondere von ihrer momentanen Auslastung abhängt, und einfach genug ist, um im Rahmen eines Charging-und-Accounting-Tools angewandt zu werden.

KAPITEL 8

# *Auktionsbasierte Preismodelle für Multiprovider-Szenarien*

## 8.1 Einführung

Nachdem in Kapitel 7 mit den dort behandelten Preisfunktionen der eine denkbare Ansatz zur Ermittlung eines dynamischen marktgerechten Preises für eine Internet-Verbindung verfolgt wurde, wobei der Schwerpunkt auf der geeigneten Modellierung einer einzelnen Teilverbindungen und dem dort vorliegenden Zusammenhang zwischen Auslastung und Preis lag, beschäftigt sich dieses Kapitel mit Preisbildung aufgrund der sogenannten "User Competition", d.h. dem direkten Wettbewerb zwischen den einzelnen Nutzern (im Gegensatz zur "Provider Competition", vgl. [Fos99]). Seinen klassischen Ausdruck findet dieser zweite Ansatz im Konzept der Auktionen. Daher beschäftigt sich vorliegendes Kapitel zunächst mit der formalen Definition von Auktionen und stellt einige verbreitete Auktionsmechanismen vor. Nachdem in Abschnitt 8.3 stichwortartig die wesentlichen Anforderungen an einen Auktionsmechanismus für Multiprovider-Szenarien zusammengefaßt werden, gehen die Abschnitte 8.4 und 8.5 auf zwei Ansätze ein, die diese Anforderungen Schritt für Schritt berücksichtigen: Während die "Delta-Auktionen" von [FSVP98] vor allem im Hinblick auf eine gleichmäßigere Verteilung des Signalisierungsverkehrs vorgeschlagen wurden, stellt das neu entwickelte "CHiPS" (Connection-Holder-is-Preferred-Scheme) [RFS99] die Belange derjenigen Nutzer in den Vordergrund, die bereits eine Verbindung über mehrere Provider aufgebaut haben und verhindern wollen, daß sie aufgrund lokal begrenzter Marktturbulenzen zusammenbricht. Abschnitt 8.6 führt dann kurz in die Simulationsumgebung "FlowSim" [Schw99] und ihre Weiterentwicklung in [Sp99] ein, die die Simulation von CHiPS erlaubt, bevor einige Simulationsergebnisse präsentiert werden.

## 8.2 Auktionsmechanismen und Utility-Funktion

### 8.2.1 Formale Definition einer Auktion

In der wirtschaftswissenschaftlichen Literatur versteht man unter einer Auktion gemeinhin "eine Institution des Markts mit einer expliziten Menge von Regeln, die die Zuteilung von Ressourcen und Preisen auf der Basis von Geboten der Marktteilnehmer bestimmt" [MM87]. Eine etwas formalere Definition findet sich in [LS97], [LS98]. Demzufolge sei gegeben eine Quantität $Q$ an *Ressourcen* und eine Menge $\Im = \{1, \ldots, I\}$ von *Spielern*, von denen jeder ein *Gebot* $s_i = (q_i, p_i) \in S_i = [0, Q] \times [0, \infty)$ über die Quantität $q_i$ zum Einheitspreis $p_i$ abgibt. Bezeichnet dann noch $S = \prod S_i$, dann ist eine Auktion zunächst definiert durch die *Allokationsregel*

$$A: \begin{cases} S \to S \\ s = (qs, ps) \to A(s) = (qA(s), pA(s)) \end{cases} \tag{8.1}$$

derzufolge Spieler $i$ die Quantität $qA_i(s)$ zum Preis $pA_i(s)$ per Ressourceneinheit zugeteilt bekommt, wobei $q$ und $p$ Operatoren darstellen, die auf ein 2-tupel angewandt jeweils die erste bzw. zweite Stelle des 2-tupels extrahieren (d.h. $qs_i = q_i$ und $ps_i = p_i$).

Die Allokationsregel $A$ heißt "zulässig", wenn gilt:

$$\begin{gathered} \sum qA_i(s) \le Q \\ A(s) \le s \end{gathered} \tag{8.2}$$

d.h. es werden nicht mehr Ressourcen verteilt als vorhanden sind, und kein Spieler erhält bzw. bezahlt mehr als sein Gebot.

Ein weiteres wichtiges Element in der formalen Definition ist die *Utility-Funktion* der einzelnen Spieler, die ihre Präferenzen ausdrückt:

$$u_i: \begin{cases} S \to [0, \infty) \\ s \to u_i(s) \end{cases} . \tag{8.3}$$

Üblicherweise setzt man für die Utility-Funktion voraus, daß sie monoton steigend, strikt konkav und stetig ableitbar ist [KMT98] (in Anlehnung an [She95] wird Verkehr, der zu solch einer Utility-Funktion führt, als "elastisch" bezeichnet [Kel97], weil in diesem Fall die Applikationen in der Lage sind, ihre Übertragungsraten an die gerade verfügbaren Ressourcen anzupassen).

Spieltheoretisch ist dann eine Auktion beschrieben durch das n-Tupel $(Q, u_1, \ldots, u_I, A)$.

### 8.2.2 Nutzerpräferenzen, Utility-Funktion und Incentive Compatibility

Wie lassen sich nun die Präferenzen von Spieler *i* genauer fassen, oder genauer gefragt: Wie sieht eine Utility-Funktion nach (8.3) konkret aus? Setzen wir voraus, daß jede Einheit der Ressource, die Spieler *i* bekommt, für ihn einen *Wert* $\vartheta_i$ besitzt. Dann läßt sich $u_i$ definieren als

$$u_i(s) = \vartheta_i qA_i(s) - qA_i(s) \cdot pA_i(s), \tag{8.4}$$

d.h. als Differenz zwischen dem, was der Spieler durch Zuteilung von $qA_i(s)$ wertmäßig erhält, und dem Betrag, den er dafür zu bezahlen hat.

(8.4) ist ein einfaches Beispiel für eine Utility-Funktion, die linear in $qA_i(s)$ ist. Wie erwähnt werden allgemeinere Utility-Funktionen in der Regel als konkav vorausgesetzt, d.h. die Zuteilung der doppelten Quantität beispielsweise hat höchstens den doppelten Nutzen für den Spieler, evtl. aber auch einen etwas darunterliegenden Wert.

Die Nützlichkeit der Utility-Funktionen ergibt sich aus der Tatsache, daß das Wissen um die privaten Präferenzen der Nutzer hinsichtlich Ressourcen und Preisen es ermöglicht, Zuteilungen durchzuführen, die “sozial erwünschter” sind als andere [WWWM98]. Genauer gesagt wird die Effizienz eines Schemas zur Ressourcenaufteilung durch das Erreichen eines Operationspunktes definiert, der eine gegebene *globale (soziale) Kostenfunktion J* minimiert; je höher also *J* ist, desto ineffizienter arbeitet das Netz. Setzt man

$$J(qA(s)) = \sum J_i(qA_i(s)) \tag{8.5}$$

als Linearkombination der Kostenfunktionen der einzelnen Nutzer an und weiterhin voraus, daß die $J_i$ kontinuierlich, monoton steigend und strikt konvex sowie stetig ableitbar sind, so gibt es einen eindeutig bestimmten Punkt, der *J* minimiert: das sogenannte *Netzwerk-Optimum* [KO99]. Das Ziel eines Netzwerkmanagers ist es dann, eine Preisstrategie anzuwenden, die zu diesem Optimum führt. Jede solche Strategie wird als “*incentive compatible*” bezeichnet.

Neben der sozialen Optimierung liegt aber auch die maximale Ausnutzung der vorhandenen Ressourcen im Interesse der Netzbetreiber. Sollte es also eine Möglichkeit geben, einzelnen Nutzern mehr Ressourcen zuzuteilen, ohne daß dies auf Kosten anderer geht, dann sollte dies gemacht werden. Eine Lösung, die in dieser Hinsicht nicht mehr zu verbessern ist, heißt “*Pareto-optimal*” [WWWM98].

Nach dem Gesagten ergibt sich als ein wesentlicher Gesichtspunkt bei der Suche nach einer Preisstrategie, daß der Netzbetreiber den Nutzern alle Informationen entlocken kann, die er für die Berechnung einer optimalen Ressourcenverteilung braucht. Insbesondere ermöglicht ihm die Kenntnis der korrekten Utility-Funktionen aller Nutzer das Auffinden einer Netzwerk-opti-

malen wie Pareto-optimalen Lösung [WWWM98]. Hieraus erklärt sich also das weitverbreitete Bestreben, ein incentive-kompatibles Auktionsschema als Preismodell für die Internet-Tarifierung zu verwenden, wie es insbesondere in der "Verallgemeinerten Vickrey-Auktion" bzw. "Second-Price Auktion" vorliegt (vgl. insbesondere [MV93], [M-M97], [LS97]).

### 8.2.3 Exkurs: Nutzerpräferenzen im Trader eines verteilten Systems

An dieser Stelle kann es instruktiv sein, zur weiteren Veranschaulichung kurz auf eine andere Anwendung des Konzepts der Utility-Funktionen einzugehen, wie sie in mehrdimensionaler Form unter dem Namen "QoS-Funktionen" in [RLT97] eingeführt wurden, um den Trading-Service unter CORBA zu flexibilisieren. Allerdings würde uns die detaillierte Vorstellung der hier entwickelten Ansätze zu weit vom Thema abbringen, deshalb sei hier kurzerhand auf Anhang B verwiesen, der die diesbezüglichen Ergebnisse zusammenfaßt und insbesondere auch auf Mechanismen zur konkreten Bestimmung von Utility-Funktionen durch den Nutzer eingeht.

### 8.2.4 Klassische Auktionsmechanismen

Im folgenden werden in Anlehnung an [WWWM98] kurz einige gebräuchliche Auktionsmechanismen vorgestellt und eingeordnet. Allgemein läßt sich festhalten, daß ein "Auktionsprotokoll" stets folgendermaßen funktioniert:

- Nutzer senden Gebote zu einem Auktionator und bekunden damit ihren Willen, Ressourcen zu erwerben und Geld dafür zu bezahlen.
- Der Auktionator *kann* eine Übersicht über die momentane Preissituation veröffentlichen und allen Nutzern zugänglich machen.

  Diese beiden Schritte können iterativ wiederholt werden!
- Schließlich legt der Auktionator eine Ressourcenverteilung fest und benachrichtigt die Nutzer darüber, was ihnen zu welchem Preis zugeteilt worden ist.

Ferner betrachten wir hier nur dezentrale Auktionsmechanismen, d.h. solche, bei denen jeder Nutzer basierend auf lokal verfügbaren Informationen seine individuelle Gebotsstrategie berechnet.

#### Ascending Auction

Dieser Auktionsmechanismus bestimmt den Preis für einzelne diskrete Güter. Die Nutzer geben sukzessive steigende Gebote ab, die sich vom Vorgebot um einen Mindestbetrag unterscheiden und vom Auktionator sofort öffentlich bekanntgegeben werden. Gehen keine Gebote mehr ein, so wird der Zuschlag an den höchsten Bieter erteilt und er erhält das Gut zu dem Preis, den er dafür geboten hat.

Die Gebotsstrategie fällt hier sehr einfach aus: Der Nutzer bietet mit, solange der gebotene Preis unter dem für ihn relevanten Wert des Gutes (vergleichbar $\vartheta_i$) liegt. Hätte er mehr zu bezahlen als ihm das Gut wert ist, steigt er aus der Auktion aus und ist darin unabhängig vom Verhalten seiner Konkurrenten.

Was die Frage nach der Optimalität dieses Auktionsschemas betrifft, so bleibt festzuhalten, daß es sich sinnvoll verhält, solange es sich jeweils um ein einzelnes Versteigerungsgut handelt. Im Fall von mehreren gleichzeitig versteigerten Gütern (wie z.B. mehreren Kanälen einer Netzwerkverbindung) kann sich das Resultat jedoch beliebig weit vom Optimum entfernen (vgl. die entsprechenden Beispiele in [WWWM98]).

### Combinatorial Auction

Um diesen Fall mehrerer gleichzeitig versteigerter Güter besser in den Griff zu bekommen, wurden verschiedene Ansätze von “kombinatorischen Auktionen” vorgeschlagen (vgl. z.B. [RSB82], [RPH97], [KS98], [HM99]). Hierbei wird ausgehend von der Menge der zur Disposition stehenden “Basisgüter” eine verallgemeinerte Menge von “Marktgütern” definiert. Zuteilungen und Preise können dann auf verschieden Weise bestimmt werden, üblicherweise als Funktion der Gebote für alle denkbaren Kombinationen von Basisgütern. Daraus ergibt sich natürlich sofort die erhöhte Komplexität derartiger Mechanismen. Immerhin lassen sich auf diese Weise Gleichgewichte erreichen, die zumindest suboptimal ausfallen.

### Generalized Vickrey Auctions

Während kombinatorische Auktionen den Problemen zu einfach gebauter Märkte durch Erhöhung der Komplexität begegnen, ist es Ziel der “Direct Revelation-Mechanismen”, die Nutzer dazu zu bringen, ihre private Information über den Wert der Güter ehrlich zu offenbaren und hieraus dann eine optimale Ressourcenverteilung zu bestimmen. Prominentestes Beispiel hierfür ist die “Verallgemeinerte Vickrey-Auktion”. Nehmen wir hierzu an, daß $\vartheta_i$ den tatsächlichen Wert eines Gutes für Nutzer *i* bezeichne. Was er tatsächlich bietet, nämlich $\vartheta_i^*$, muß nicht unbedingt gleich $\vartheta_i$ sein - der Nutzer kann ja auch unehrlich sein, weil er sich von einem abweichenden Gebot einen Vorteil verspricht. In diesem Fall ist es für den Auktionator aber nicht möglich, eine optimale Lösung zu berechnen.

Der Clou der Verallgemeinerten Vickrey-Auktion besteht nun darin, daß erfolgreiche Bieter als Preis nicht ihr Gebot bezahlen, sondern das Gebot desjenigen Konkurrenten, der gerade nicht mehr erfolgreich war (in gewissem Sinn also den Preis, zu dem der Markt geräumt wird). Im Fall von teilbaren Gütern wird dieser Mechanismus zur “Second Price-Auktion” [LS97], bei der sich der zu entrichtende Preis daraus ergibt, was von den anderen Mitbietern gezahlt worden wäre, wenn das eigene Gebot gar nicht existiert hätte.

Es läßt sich nun relativ einfach zeigen ([WWWM98], [LS97]), daß die optimale Strategie für den Bieter darin besteht, seinen tatsächlichen Wert $\vartheta_i$ als Gebot zu offenbaren, da er genau in

diesem Fall seine Utility-Funktion (8.4) maximieren kann. Dies wiederum ermöglicht es dem Auktionator, eine optimale Lösung zu berechnen, daher ist die Verallgemeinerte Vickrey-Auktion “incentive compatible”.

## 8.3 Anforderungen im Multiprovider-Szenario

In der Literatur ist es weitverbreiteter Konsens, aus den am Ende des vorigen Abschnitts vorgebrachten Gründen das Konzept der Verallgemeinerten Vickrey-Auktionen als Grundlage für auktionsbasierte Internet-Tarifierung zu verwenden. In diesem Abschnitt werden nun Anforderungen an eine Erweiterung dieses Auktionsmechanismus formuliert, wie sie sich aufgrund der Multiprovider-Sicht der betroffenen Verbindungen ergeben.

Als Grundlage hierfür wird angenommen, daß für die Signalisierung ein Reservierungsprotokoll (z.B. RSVP, vgl. Abschnitt 6.3.2) zur Verfügung steht, das in der Lage ist, Preis- und Gebotsinformationen zu übertragen. Weiterhin wird angenommen, daß alle Provider unabhängig voneinander lokale Auktionen für alle Verbindungen durchführen, die momentan auf einem bestimmten Ausgangs-Link aktiv sind.

Die Verwendung eines derartigen “Smart Market”-Schemas, das in der Lage ist, auf Veränderungen des Marktes sofort zu reagieren, zur realen Preisbestimmung für Flows im Internet wirft im Hinblick auf Multiprovider-Szenarien folgende Schwierigkeiten auf:

- Auktionen finden immer zu diskreten Zeitpunkten statt, die von ISP zu ISP verschieden sind und sich nicht ohne weiteres synchronisieren lassen.
- Daher ist es für den Nutzer schwierig, konsistente Informationen über Anzahl, Lokalisierung, aktuelle Marktsituation usw. der einzelnen Ressourcen zu erhalten, die er für den Aufbau seiner Ende-zu-Ende-Verbindung aneinanderzureihen hat.
- Bieter müssen nach Abgabe ihres Gebotes bis zur nächsten Auktion untätig bleiben, bis ihre Anfrage beantwortet werden kann.
- Bei Multiprovider-Verbindungen (d.h. Verbindungen über mehrere Knoten bzw. ISPs) trifft dies zudem nacheinander für jeden Knoten bzw. jede individuelle Teilverbindung zu, was eine durchschnittliche weitere Verzögerung des Verbindungsaufbaus in der Größenordnung des Produkts aus Knotenanzahl und mittlerer Auktionsperiode (d.h. Zeit zwischen zwei Auktionen) zur Folge hat.
- Nach einer Auktion müssen alle Auktionsteilnehmer vom Ausgang unterrichtet werden, was einen erheblichen Signalisierungs-Burst mit sich bringen kann.
- Geht bei einer Verbindung über mehrere Provider nur eine Auktion lokal verloren, so ist davon sofort die gesamte Verbindung betroffen und wird im Normalfall wohl abgebrochen.

- Auch die Frage, wie denn ein vorgegebenes Gesamt-Budget (für die komplette Verbindung) am besten auf die Gebote für die einzelnen lokalen Auktionen aufzuteilen ist, stellt sich als nicht-trivial heraus.

Die folgenden Untersuchungen widmen sich Ansätzen, die auf die genannten Problematiken eingehen. Zunächst wird das Konzept der "Delta-Auktionen" vorgestellt, das zumindest die mit der Signalisierung verbundenen Schwierigkeiten angeht. Danach präsentieren wir in Abschnitt 8.5 ein neues Auktionsschema, das speziell für Multiprovider-Verbindungen entwikkelt wurde. Abschnitt 8.6 geht auf die Weiterentwicklung einer flow-basierten Simulationsumgebung ein, die eine simulative Bewertung des neuen Schemas ermöglicht, und stellt ausgewählte Simulationsergebnisse vor, bevor Abschnitt 8.7 in einem Ausblick die mögliche Einbindung der vorgestellten dynamischen Preismodelle in ein reales Charging- und Accounting-Tool beschreibt, wie es im Rahmen des CATI-Projektes derzeit entwickelt wird.

## 8.4 Delta-Auktionen

Ein erster Ansatz zur Bewältigung der geschilderten Problematiken wurde mit Hilfe der sogenannten "Delta-Auktionen" versucht ([FSVP98], [RFS99], [SRL99]). Wir haben gesehen, daß die fehlende Synchronisation zwischen den einzelnen lokalen Auktionen einer Multiprovider-Verbindung zu einer Verzögerung von durchschnittlich $(n/2)\sum P_i$ führt, wobei $n$ die Anzahl ISPs und $P_i$ das Auktionsintervall (bzw. die Reservationsperiode) beim Provider $i$ bezeichnet. Da die sogenannten Auffrischungsperioden, nach denen eine Reservation normalerweise wiederholt bzw. bestätigt werden muß, bei den im Internet gebräuchlichen Reservierungsprotokollen in der Größenordnung von zehn Sekunden liegen, ist eine unmodifizierte Anwendung des Auktionsgedankens nicht praktikabel.

Deshalb wurde in [FSVP98] das Konzept der Delta-Auktionen vorgeschlagen, das ein kontinuierliches Stattfinden des Auktionsvorganges unterstützt. Die Grundidee besteht hierbei darin, ankommende Anfragen unmittelbar weiterzuverarbeiten und Gebote, die offensichtlich zu niedrig ausgefallen sind, von vorneherein zurückzuweisen. Damit wird verhindert, daß die betroffenen Nutzer untätig auf den Auktionsausgang bei den einzelnen Providern warten, der ja mit Sicherheit negativ ausfallen wird. Ausreichend hohe Gebote hingegen werden provisorisch akzeptiert, wobei allerdings ausdrücklich der Vorbehalt besteht, daß evtl. später ankommende Gebote eine solche vorläufige Zusage wieder hinfällig machen.

Hauptvorteil dieses Schemas ist, daß die Nutzer sehr früh über die Zurückweisung einer Reservation informiert werden, ferner wird dadurch der Signalisierungsverkehr ziemlich gleichmäßig über die Zeit verteilt. Die provisorische Zusage bei entsprechend hohem Gebot hat für den Nutzer zumindest den Charakter einer positiven Rückmeldung, auch wenn eine gewisse Unsi-

cherheit bleibt, ob die Reservierung dann (nach Ablauf der Auktionsperiode) auch wirklich bestätigt werden kann.

Delta-Auktionen wurden unter zwei Reservationsprotokollen implementiert und evaluiert [Fos99], nämlich SSP (State Setup Protocol) [APY98] und RSVP (Resource Reservation Protocol [ZDE+93], [BZB+97]), die beide empfängerorientiert arbeiten. An dieser Stelle soll jedoch nur kurz auf die bei Verwendung von RSVP erhaltenen Resultate eingegangen werden [RFS99].

Insbesondere wurde das Verhalten von Delta-Auktionen unter RSVP durch Simulationen mit der ns-2-Simulationsplattform [ns-2] untersucht. Um weiteren Signalisierungsverkehr zu sparen, wurden hierbei die Gebote in Form eines sog. “Bid-Faktors” ausgedrückt, d.h. relativ zum momentanen Marktpreis. Ein Bid-Faktor von 1.1 etwa entspricht einem Gebot in Höhe von 110% des derzeitigen Marktpreises. Die Auswertung der Simulationen hat allerdings ergeben, daß dies zu einer Benachteiligung von Geboten führen kann, die relativ früh innerhalb einer Auktionsperiode beim Auktionator eintreffen. Daher wurde als weitere Möglichkeit untersucht, den Marktpreis der vorangegangenen Auktion als fixe Grundlage für ein relatives Gebot (wiederum über einen entsprechenden Bid-Faktor) zu verwenden.

Abbildung 8-1 zeigt das Ergebnis einer Simulation für acht Bieter, die alle die gleiche Verbindung benutzen wollen und daher innerhalb derselben Auktion miteinander konkurrieren. Bieter 1 gibt sein Gebot stets als erster ab, während Bieter 8 stets der letzte ist. Die durchgezogene Linie stellt die Anzahl der von Bieter 1 bis 8 gewonnenen Auktionen dar und entspricht dabei der Gebotsberechnung relativ zum aktuellen Marktpreis.

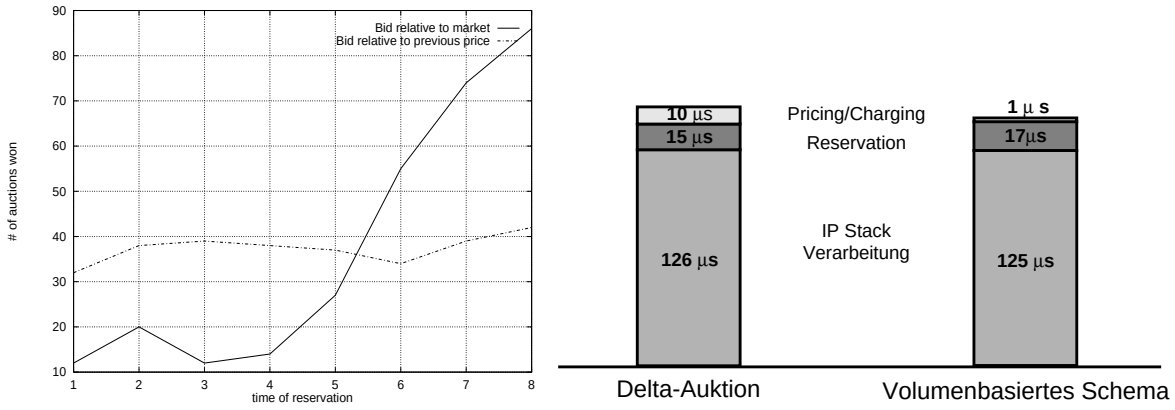


**Abbildung 8-1:** “Wer zu spät kommt, den belohnt das Leben” (frei nach [Gor99]), jedenfalls bei Delta-Auktionen: Simulationsergebnis unter ns-2 (links). Rechts Vergleich der Zeiten für Protokolloverhead zwischen Delta-Auktionen und einem volumenbasierten Preisschema (aus [SRL99]).

In den 300 durchgeführten Auktionen erzielten die Zuspätkommenden einen klaren Vorteil, da ihre Gebote auf den Preisinformationen des kurz vor der Schließung stehenden Marktes

basiert. Im Gegensatz hierzu basiert die gestrichelte Linie auf der Gebotsberechnung relativ zum Marktpreis der vorangegangenen Periode und ergibt eine relativ gleichmäßige Verteilung der Anzahl gewonnener Auktionen über die einzelnen Bieter. Der Nachteil letzterer Variante besteht freilich darin, daß es sich dabei genau genommen nicht mehr um eine reine Second-Price-Auktion handelt, wobei sich die Abweichung aber im Rahmen der Marktfluktuation bewegt.

Ein weiterer erwähnenswerter Aspekt, der simulativ untersucht wurde, betrifft die Frage, ob ein Nutzer durch geschickte Analyse von Signalisierungsdaten die genauen Auktionszeitpunkte herausfinden kann, um auf diese Weise einen Wettbewerbsvorteil zu gewinnen. Es hat sich jedoch herausgestellt, daß ein derartiges Fairneßproblem im Fall einer genügend hohen Bieter- und Provideranzahl nicht auftritt. In der Praxis ist das Abfangen von Signalen zwischen zwei Hosts, zu denen der Angreifer keinen Zugang hat, zwar nicht unmöglich, aber in der Regel auch nicht gerade einfach. Durch die Vielzahl von betroffenen Verbindungen und Flows wird die Erlangung eines Marktvorteils noch weiter erschwert. Schließlich wurden auch noch die für Delta-Auktionen benötigte Rechenzeit mit der eines normalen volumenbasierten Schemas für den Fall einer einzelnen Reservierung eines IP-Telefongesprächs verglichen (Abbildung 8-1 rechts). Während die Verarbeitungszeit für den IP-Stack und die Reservierung dabei im wesentlichen gleich bleibt, ergibt sich der zusätzliche Pricing-Overhead typischerweise zu 0.7% für das volumenbasierte Schema bzw. 6.7% bei der Delta-Auktion [SRL99]. Für weitere Details dieser Untersuchungen sei auf [Fos99] verwiesen.

## 8.5 CHiPS: Das Connection-Holder-is-Preferred-Scheme

### 8.5.1 Grundzüge des Verfahrens

Wie in Abschnitt 8.4 erläutert, stellt das Konzept der Delta-Auktionen eine Lösung für das Verzögerungsproblem bereit, das dadurch auftritt, daß eine Vielzahl von Auktionen benötigt wird, um eine Verbindung aufzubauen, die über entsprechend viele ISPs verläuft. Im vorliegenden Abschnitt kümmern wir uns nun um einige weitere Punkte, die im Zusammenhang mit Multiprovider-Szenarien problematisch sind. Nehmen wir hierzu einmal an, eine Ende-zu-Ende-Verbindung sei bereits aufgebaut, und ihre Reservierung wird gemäß dem verwendeten Reservationsprotokoll von Zeit zu Zeit erneuert. Unglückliche Umstände können nun bewirken, daß sich die Marktsituation auf einem kurzen Abschnitt der Verbindung, z.B. innerhalb eines kleinen ISPs, so ändert, daß die entsprechende lokale Auktion plötzlich für den Nutzer verlorengeht. Dann kann dieser lokale Verlust eventuell zur Folge haben, daß die gesamte Verbindung abgebrochen wird, da die entsprechenden Ressourcen immer eine Ende-zu-Ende-Garantie besitzen müssen.

Als Ausweg schlagen wir im folgenden ein Schema vor, das versucht, solche lokalen Probleme erst einmal lokal zu lösen, d.h. im betroffenen ISP als dem Auktionator selbst. Hierzu ist es erforderlich, den Ausgang einer Auktion neu zu interpretieren, d.h. formal gesehen die in Abschnitt 8.2.1 definierte Allokationsregel zu ändern. Zwar wird hierbei die Auktion weiterhin dazu verwendet, anhand der Gebote eine Rangliste unter den Bietern herzustellen wie auch den Marktpreis festzulegen (und zwar anhand der üblichen Second-Price-Auktion). Aber für einen neu hinzukommenden Bieter, d.h. einen, der eine neue Verbindung aufbauen möchte, ist ein vorderer Platz in der Rangliste noch nicht gleichbedeutend damit, daß seine Ressourcenanforderung auch erfüllt wird. Vielmehr werden die Inhaber von bereits laufenden Verbindungen, die ja ihre Reservierung von Zeit zu Zeit erneuern müssen, bevorzugt behandelt, indem ihnen, sollten sie die Auktion aufgrund des plötzlich fluktuierenden Marktes durch ein zu niedriges Gebot verloren haben, eine zweite Chance zugeteilt wird. Hierzu wird ihre bereits bestehende Verbindung nicht abgebrochen, sondern noch eine Auktionsperiode lang aufrechterhalten. Währenddessen erhalten sie die Möglichkeit, ihr Gebot nachträglich so zu erhöhen, daß es über dem ermittelten Marktpreis liegt, und sich damit in die Reihe der Gewinner hineinzuschmuggeln.

Das resultierende Schema hat den Namen CHiPS (für “Connection-Holder-is-Preferred-Scheme”) erhalten (vgl. [RFS99]). Es wurde im Hinblick auf die in den beiden vorangegangenen Abschnitten dargestellten Aspekte und Konzepte entwickelt und weist folgende Vorteile auf:

- Es handelt sich um ein dynamisches Preismodell für Multiprovider-Verbindungen, das für den Kontext der realen Entwicklung eines Internet Charging- und Accounting-Systems unter RSVP geeignet ist.
- Sein Hauptvorteil ist die Tatsache, daß der lokale Verlust einer Teilverbindung aufgrund einer plötzlichen Änderung der Marktsituation nicht automatisch zum Abbruch der bereits bestehenden Ende-zu-Ende-Verbindung führen muß.
- Ein weiterer Vorteil ergibt sich aus der hierzu notwendigen Neuinterpretation des Nutzerbudgets: Klassische Auktionsschemata beruhen darauf, daß jeder Nutzer ein Gesamtbudget besitzt (auch “Spending Cap” genannt), welches die Höhe der Gebote begrenzt (vgl. z.B. [LS97]). Ein Nutzer kann also nicht mehr bieten, als sein Budget zuläßt. Andererseits ist es ja eine Haupteigenschaft der Verallgemeinerten Vickrey-Auktionen, daß der tatsächliche Marktpreis in der Regel unterhalb des Gebots zu liegen kommt, der Nutzer also regelmäßig weniger bezahlt als sein Budget eigentlich zuließe. Der Vorteil von CHiPS besteht nun darin, daß es gewissermaßen eine nachträgliche Umverteilung der noch freien Gelder zuläßt und somit dafür sorgt, daß das Nutzerbudget die tatsächlichen Aufwendungen begrenzt und nicht mehr die Höhe der Gebote.

Wie im Fall der Delta-Auktionen erfordert CHiPS die regelmäßige Durchführung von Auktionen für jede Einzelressource, wobei Reservationen nur für die gesamte Dauer der entsprechen-

den Auktionsperiode zulässig sind. Außerdem darf jeder Nutzer höchstens ein Gebot pro Auktionsperiode einreichen.

Natürlich ist es im allgemeinen nicht ausgeschlossen, daß zu einem bestimmten Zeitpunkt mehr als eine lokale Auktion verlorengeht (wenn das Gesamtbudget zu niedrig dimensioniert ist, kann dies sogar sehr schnell der Fall sein), aber aus Gründen der Übersichtlichkeit beschränken wir unsere Überlegungen zunächst einmal auf den Fall des Verlustes einer Einzelauktion, was eine typische Situation für den Fall eines Flaschenhalses im Netz darstellt. Die Verallgemeinerung auf den Fall mehrerer verlorener Auktionen bietet keine grundlegenden Schwierigkeiten. Darüberhinaus kann der simultane Verlust einer großen Anzahl von Auktionen ein deutlicher Hinweis an den Nutzer sein, doch einmal über eine Erhöhung seines Budgets nachzudenken oder gar seine Bemühungen für den Moment erst einmal ganz einzustellen.

### 8.5.2 Auktionen unter CHiPS

Auch wenn wir uns in einem Multiprovider-Szenario bewegen, finden die einzelnen Auktionen doch immer noch lokal statt, d. h. beispielsweise bei jedem einzelnen Provider. Deshalb wird im folgenden zunächst verdeutlicht, wie eine solche Einzelauktion unter CHiPS abläuft.

Betrachten wir hierzu eine einzelne Ressource der Kapazität $C$, für die Auktionen jeweils an den Zeitpunkten $t_1$, $t_2$ usw. stattfinden, wobei $t_2 - t_1 = t_3 - t_2 = ... = const$ ist. Nehmen wir an, zum Zeitpunkt $t_1$ ist bekannt, daß zum Zeitpunkt $t_2$ eine bestimmte Anzahl von Reservationen nicht erneuert wird, da die entsprechende Verbindung entweder freiwillig oder gewaltsam beendet wird. Deshalb ist abzusehen, daß zu diesem Zeitpunkt Kapazität in Höhe von $C'$ frei wird. Außerdem kommen in der Zeit zwischen $t_1$ und $t_2$ eine bestimmte Anzahl von neuen Reservierungsanfragen mit entsprechenden Geboten beim Provider an. In diesem Fall stellen sich zwei Fragen: Welche der Anfragen sollen durch Zuteilung entsprechender Ressourcen erfüllt werden? Und welcher Preis ist für die Nutzung von Ressourcen zu entrichten? Die klassische Verallgemeinerte Vickrey-Auktion koppelt diese beiden Fragen insofern, als die höchsten Bieter ihre Wünsche erfüllt bekommen, und zwar zu genau dem Preis, bei dem der Markt geräumt wird. Demgegenüber werden diese beiden Aspekte unter CHiPS entkoppelt, und zwar wie nachfolgend beschrieben.

#### Erfüllung von Reservationsanfragen

Die offensichtliche Antwort auf die erste Frage lautet: Nach wie vor werden die Anfragen der höchsten Bieter erfüllt, allerdings ist dies nur der Fall für die freie Kapazität $C'$ (statt für die gesamte Kapazität $C$ wie bei der Verallgemeinerten Vickrey-Auktion). Außerdem findet die Auktion hierfür nur unter den neu angekommenen Anfragen statt. Anders ausgedrückt: Eine ganz gewöhnliche Auktion unter allen zwischen $t_1$ und $t_2$ angekommenen Geboten findet statt, und zwar über die Kapazität $C'$; hierbei gewinnen wie üblich die höchsten Gebote und erhalten entsprechend Ressourcen zum Zeitpunkt $t_2$ zugewiesen. Dieses Vorgehen ermöglicht übrigens

das bereits von der Delta-Auktion her bekannte vorzeitige Versenden von negativen Bescheiden, sobald feststeht, daß ein Gebot aufgrund der Marktlage keine Chance haben wird.

#### Preisbestimmung

Um den korrekten Marktpreis zum Zeitpunkt $t_2$ zu bestimmen, muß nun eine "fiktive Auktion" unter allen betroffenen Nutzern, d.h. denen, die bereits eine Verbindung haben, und den neu hinzugekommenen Bietern, stattfinden. Unter all diesen ermittelt eine Second-Price-Auktion wie üblich den Marktpreis. Nun kann es allerdings vorkommen, daß einige der Verbindungsinhaber Gebote unterhalb dieses Marktpreises eingereicht haben (und damit eigentlich die Auktion verloren haben). Diese Nutzer erhalten umgehend eine Nachricht, die ihnen den momentanen Marktpreis mitteilt und sie auffordert, im Nachhinein ihr Gebot entsprechend anzupassen, widrigenfalls ihre Verbindung nach Ablauf der nächsten Auktionsperiode gewaltsam abgebrochen wird. Entschließt sich der Nutzer zur Gebotserhöhung, so hat er für die laufende Auktionsperiode entsprechend mehr zu entrichten, verweigert er die Erhöhung, so bezahlt er lediglich sein bisheriges Gebot - die Differenz zum Marktpreis hat der ISP zu tragen. Auf diese Weise ist sichergestellt, daß ein Nutzer niemals einen höheren Betrag bezahlen muß, als er vorher zugestimmt hat.

### 8.5.3 Konsequenzen

Aus dem bisher Gesagten lassen sich die folgenden wichtigen Konsequenzen ziehen:

- Da alle vorliegenden Gebote in der erwähnten fiktiven Auktion berücksichtigt werden, entspricht der dadurch ermittelte Preis dem korrekten Marktpreis einer Second-Price-Auktion (abgesehen von den Fällen, in denen ein Verbindungsinhaber die fällige Gebotserhöhung ablehnt, weil er z.B. sowieso gleich fertig ist, und der ISP das Risiko tragen muß. Letzterer kann sich dagegen z.B. absichern, indem er eine kleine allgemeine Gebühr für die Deckung seiner Grunddienstleistungen erhebt, etwa in Form einer flat fee oder eines Aufschlages auf den jeweiligen Marktpreis). Aus diesem Grund überträgt sich auch die Eigenschaft der "incentive compatibility" von den Second-Price-Auktionen auf CHiPS, wie simulativ in Abschnitt 8.6.3 demonstriert wird.
- Kein Nutzer muß jemals mehr zahlen, als er vorher zugestimmt hat. Falls sich der Preis für eine Verbindung erhöht, kann er jederzeit ohne finanzielle Folgen zurücktreten.
- Bei klassischen Auktionen konzentriert sich, wie oben ausgeführt, die Signalisierungslast meist stark um die Auktionszeitpunkte. CHiPS trägt zur gleichmäßigen zeitlichen Verteilung der Signalisierungslast bei. Beispielsweise kann die erste Hälfte einer Auktionsperiode dazu genutzt werden, die für evtl. erforderlich nachträgliche Gebotserhöhungen notwendigen Nachrichten auszutauschen, während die zweite Hälfte der Auktionsperiode vor allem dazu verwendet wird, vorzeitige Absagen an die Nutzer zu senden, die bei der nächsten Auktion keine Chance haben.

- Wichtigster Vorzug des neuen Schemas ist die Tatsache, daß es laufende Verbindungen nicht unterbricht. Im Fall einer plötzlichen gravierenden Marktfluktuation bleibt stets mindestens eine Auktionsperiode lang Zeit für Gegenmaßnahmen.
- Das vorgeschlagene Schema ist äquivalent zu einer Second-Price-Auktion, bei der die Menge der Gewinner nicht identisch ist mit der Menge derjenigen, deren Anforderungen erfüllt werden. Aber der Marktpreis selbst wird korrekt ermittelt, da die fiktive Auktion alle mitbietenden Nutzer in bester "Smart Market"-Tradition berücksichtigt.
- Außerdem ist es jetzt möglich, weiteren Signalisierungsverkehr dadurch einzusparen, daß man Nutzer, deren Verbindung steht, mit ihren bisherigen Geboten quasi "stillschweigend" weiter mitbieten läßt (nachdem sich für sie zunächst sowieso kein Grund ergibt, ihr einmal erfolgreiches Gebot zu ändern).

Zusammenfassend kann man festhalten, daß das vorgeschlagene Schema die in Abschnitt Abschnitt 8.3 aufgelisteten Probleme von Auktionen im Multiprovider-Szenario elegant umgeht. Sobald einmal eine Verbindung aufgebaut ist, kann der Nutzer jederzeit einmal eine Auktion verlieren, ohne daß die Verbindung notwendigerweise unterbrochen würde. Stattdessen erhält der Nutzer in diesem Fall eine Aufforderung, seine Gebühr der veränderten Marktsituation anzupassen. Daraufhin kann er bequem sein Gesamtbudget zu Rate ziehen und gegebenenfalls den neuen Preis für die Teilverbindung akzeptieren. Insgesamt führt dies letztlich zur Entkopplung der einzelnen zu einer bestimmten Verbindung gehörigen Auktionen.

### 8.5.4 Implementationsaspekte

In diesem Abschnitt werden einige implementationsspezifische Aspekte näher beleuchtet. Diese reichen von der Frage, wie denn im Fall einer Multiprovider-Verbindung aus dem vorhandenen Gesamtbudget die für die einzelnen Auktionen getrennt abzugebenden Gebote bestimmt werden können, über eine Bewertung der Brauchbarkeit von RSVP als zugrundeliegendes Protokoll bis zu zwei Vorschlägen zur konkreten Durchführung einer nachträglichen Gebotserhöhung im Fall einer verlorenen Auktion.

#### Gebote im Multiprovider-Szenario

Wir nehmen also wiederum an, daß dem Nutzer ein Gesamtbudget $B$ zur Verfügung steht, das er maximal für das Zustandekommen einer Multiprovider-Verbindung auszugeben gewillt ist. Andererseits sind hierfür Gebote für eine Anzahl $n$ unabhängiger Auktionen abzugeben, und die Frage nach der Aufteilung von $B$ auf die einzelnen Gebote stellt sich. Denkbar wäre es z.B., bei jeder Auktion gleichviel zu bieten. Diese Strategie läßt allerdings unberücksichtigt, daß in der Regel nicht alle Teilstücke einer Multiprovider-Verbindung gleichermaßen verstopft sind. Üblicherweise ist bei Transatlantikverbindungen beispielsweise das Teilstück zwischen Europa und den USA überlastet, während innerhalb Europas bzw. der Vereinigten Staaten die

Kapazitäten bei weitem ausreichen (was den dafür zu entrichtenden Marktpreis in die Nähe von Null drückt).

Aus diesem Grund wird von folgendem adaptiven Schema ausgegangen [Sp99]:

- Zunächst ist die Summe der aktuellen Marktpreise $m_i$, $i = 1, ..., n$, zu ermitteln:

$$S = \sum_{i=1}^{n} m_i \tag{8.6}$$

- Sodann wird der sog. *Bid-Faktor* $f$ als Quotient von Spending Cap und gesamtem Marktpreis berechnet:

$$f = \frac{B}{S} \tag{8.7}$$

Der Bid-Faktor gibt (wie bereits von den Delta-Auktionen her bekannt) nicht den absoluten Wert des abgegebenen Gebotes an, sondern wird an alle $n$ Auktionatoren übermittelt, wo er dann wiederum mit dem dortigen Marktpreis multipliziert wird, um die absolute Höhe des Gebotes zu ergeben. Dadurch erlaubt er eine Gewichtung der Gebote in Anlehnung an die aktuelle Marktsituation entlang der Multiprovider-Verbindung.

Aus der Verwendung dieses adaptiven Schemas ergibt sich hinsichtlich der Signalisierung sogleich die Abfolge von “Quote Phase” und “Bid Phase”, wie sie im nächsten Abschnitt erläutert wird.

### Implementierung in den RSVP-Algorithmus

Zur Realisierung des adaptiven Bidfaktor-Schemas zur Gebotsberechnung ist die Durchführung einer eigenen “Quote Phase” zur Ermittlung des aktuellen Gesamtpreises nach (8.6) erforderlich, bevor in der “Bid Phase” die eigentliche Gebotsabgabe erfolgt (vgl. Abbildung 8-2).

#### *Quote Phase*

Der Sender schickt eine ganz normale PATH-Nachricht zum Empfänger. Auf dem Rückweg benötigt die RESV-Nachricht ein zusätzliches Objekt, das “SUM Object”, welches den aktuellen Gesamtpreis für den entsprechenden Pfad ermittelt, und zwar durch sukzessives Aufaddieren der Marktpreise für die einzelnen Teilstücke. Sobald die Nachricht zum Sender zurückgekehrt ist, kann daraus dann gem. (8.7) der Bid-Faktor berechnet werden.

#### *Bid Phase*

Auch diese Phase beginnt mit dem Senden einer *PATH*-Nachricht durch den Sender (was unter Standard-RSVP alle 30 Sekunden wiederholt werden muß, um die Reservationen aufzufri-

schen). Der berechnete Bid-Faktor wird dieser Nachricht mitgegeben, wobei sogar das eben eingeführte SUM Object hierfür wiederverwendet werden kann. Der Bid-Faktor wird daraufhin (neben den üblichen Parametern wie Flow-Id, vorhergehender Hop und Empfänger des Flows) in jedem Hop gespeichert. Nach der Ankunft der *PATH*-Nachricht beim Empfänger schickt dieser eine *RESV*-Nachricht zurück, wobei das SUM Object nun wiederum für das Einsammeln der (inzwischen aktualisierten) Marktpreise zuständig ist.

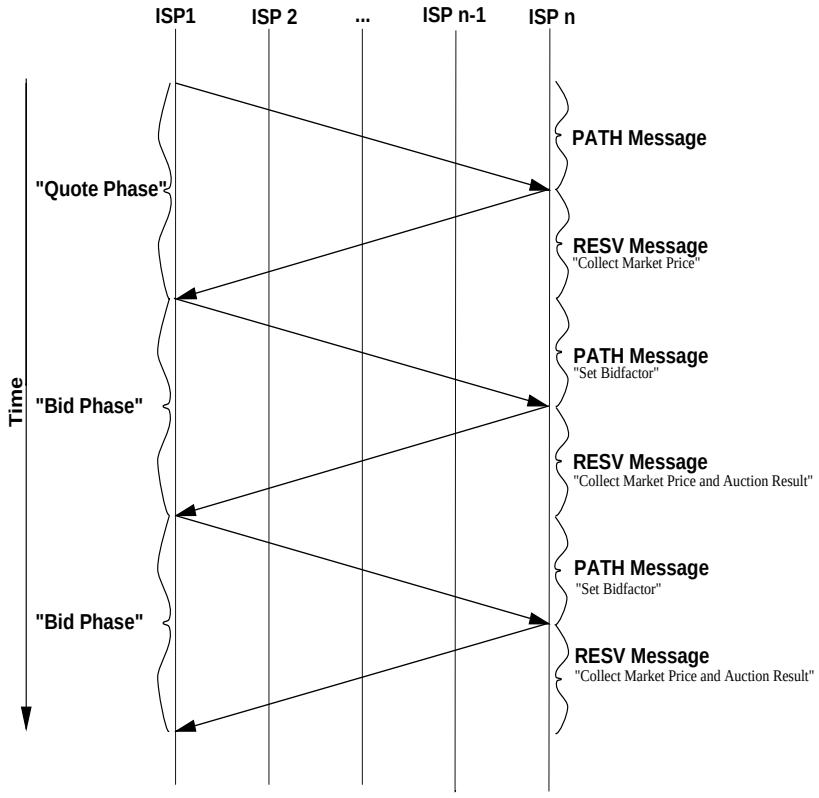


**Abbildung 8-2:** Message Sequence Chart für CHiPS-Auktionen

Bislang sind mit Ausnahme des neudefinierten Objekts noch keine fundamentalen Änderungen zum Standard-RSVP zu verzeichnen. Die letztgenannte *RESV*-Nachricht muß nun allerdings auch noch die Auktionsergebnisse transportieren. Hierzu ist ein zweites neues Objekt, das "WON_AUCTION Object", vorzusehen. Die *RESV*-Nachricht wartet also an jedem Hop, bis das Auktionsergebnis vorliegt, sammelt den aktuellen Marktpreis ein und geht zum nächsten Hop. Wieder beim Sender angekommen, wird ein neuer Bid-Faktor berechnet und das entsprechende Objekt der nächsten *PATH*-Nachricht aktualisiert, bevor der Zyklus mit der nächsten Bid Phase fortgesetzt wird, und zwar solange, bis eine Auktion verloren geht oder die Datenübertragung von Nutzerseite beendet wird.

### Vorgehen bei nachträglicher Gebotserhöhung

Zu klären bleiben jetzt vornehmlich noch die Frage, wie denn eine fällige nachträgliche Gebotserhöhung in der Praxis realisiert werden kann. Hierfür werden im folgenden zwei Ansätze vorgeschlagen.

#### *Direkte Benachrichtigung*

Unter RSVP ist es möglich, Nachrichten direkt an einen individuellen ISP zu adressieren. Wir haben bereits im einleitenden Abschnitt gesehen, daß sich RSVP-Nachrichten zur Übertragung von Geboten verwenden lassen. Dies soll hier noch etwas genauer ausgeführt werden.

RSVP-Nachrichten bestehen aus einem Header und einem Body, der seinerseits eine Menge von RSVP-Objekten enthält (siehe Abschnitt 6.3.2). Zusätzlich zu den Standard-RSVP-Objekten wurden eine Anzahl von sogenannten “Pricing Objects” entwickelt (vgl. [Fos99]), die ebenfalls im RSVP-Nachrichtenbody transportiert werden können. Unter anderem findet sich unter diesen neuen Objekten ein “BID Object”, das den oben erläuterten Bid-Faktor (relativ zum aktuellen oder vorherigen Marktpreis) beinhalten kann. Das “PROVIDER Object” enthält einen Schlüssel, der es erlaubt, den ISP zu identifizieren, bei dem die entsprechende lokale Auktion verlorengegangen ist. Folglich kann man unmittelbar nach dem Verlust der Auktion dem Nutzer eine RSVP-Fehlermeldung schicken, die den spezifischen Provider-Schlüssel enthält. Der Nutzer kann daraufhin explizit das Gebot für die betreffende Auktion erhöhen.

#### *Das AMF-Schema*

Unser zweiter Vorschlag vermeidet das Senden einer Nachricht vom Nutzer an einen bestimmten ISP, der irgendwo auf halber Strecke der Verbindung liegen kann, indem es die Gebotsstruktur modifiziert. Es wird sich herausstellen, daß dies noch einen weiteren, sehr positiven Nebeneffekt mit sich bringt.

Nehmen wir hierzu an, daß von nun an ein Gebot $b_i$ nicht länger die Form “aktueller Marktpreis mal Bid-Faktor” besitzt, sondern noch einen zusätzlichen konstanten Term aufweist:

$$b_i = a_i + m_i \cdot f \tag{8.8}$$

wobei $m_i$ den aktuellen Marktpreis beim ISP $i$ bezeichnet, $f$ den im “BID Object” transportierten Bid-Faktor und $a_i$ eine Konstante, die lokal beim ISP als Auktionator gespeichert ist und zu Null initialisiert wird[1].

Sobald nun eine Auktion verlorengeht, wird eine RSVP-Fehlermeldung vom entsprechenden ISP an beide Endpunkte der Verbindung geschickt. In dieser Meldung ist die Differenz $d$ zwischen dem aktuellen Marktpreis und dem (zu niedrigen) Gebot des Nutzers enthalten. Auf dem

1. Der Name “AMF-Schema” wurde übrigens mangels einer besseren Idee in Anlehnung an die Form von (8.8) gewählt.

Weg zu Sender bzw. Empfänger der betroffenen Verbindung passiert die Fehlermeldung alle an dieser Verbindung beteiligten ISPs und fordert jeden dazu auf, seine entsprechende Konstante $a_i$ um den Wert von $d$ zu erniedrigen. Jetzt ist der Nutzer aufgefordert, zu entscheiden, ob sein Gesamtbudget eine Steigerung des fraglichen Gebotes um $d$ erlaubt. Ist dies der Fall, so verwendet der Nutzer eine normale RSVP-Ende-zu-Ende-Nachricht, die jeden ISP auffordert, die entsprechende Konstante $a_i$ wieder um den Wert $d$ zu erhöhen. Auf diese Weise ändern sich die Gebote bei den gewonnenen Auktionen letztlich nicht, wohingegen bei dem ISP mit der verlorenen Auktion das Gebot über die Konstante $a_i$ schließlich um genau den Wert $d$ vergrößert worden ist, um den es ursprünglich zu gering ausgefallen war.

Der weitere Vorteil der Gebotsstruktur von (8.8) liegt darin, daß hiermit die Gebote für besonders riskante Auktionen, d.h. solche mit stark schwankenden Marktsituationen, automatisch höher ausfallen als für Auktionen, bei denen sich der Marktpreis kaum je ändert. Das liegt daran, daß der einmal nach dem Auktionsverlust erhöhte Wert von $a_i$ für den Rest der Dauer der Verbindung erhöht bleibt. Anders ausgedrückt wird also bereits vor Ermittlung des Bid-Faktors, mit dem jeweils die aktuellen Marktpreise zu multiplizieren sind, ein bestimmter Teil des Gesamtbudgets vorab an strategisch wichtigen Auktionen plaziert. Aufgrund dieses Vorteils wurde das AMF-Schema auch den Simulationen des nun folgenden Abschnitts zugrundegelegt.

## 8.6 Simulation von Multiprovider-Szenarien unter FlowSim - Umgebung und Resultate

Zum Abschluß des Kapitels wird in diesem Abschnitt eine erste simulative Untersuchung und Bewertung des vorgeschlagenen Ansatzes vorgenommen. Hierzu wird kurz auf die verwendete Simulationsumgebung eingegangen und das Simulationsszenario spezifiziert, bevor einige ausgewählte Resultate vorgestellt werden.

### 8.6.1 Die FlowSim-Umgebung

Das Java-basierte Simulationstool FlowSim wurde von [Schw99] ursprünglich zur Simulation von Mechanismen zum SLA-Trading ([FSP99], vgl. auch Abschnitt 6.3.1) entwickelt und von [Sp99] um Klassen zur Simulation von Auktionsmechanismen erweitert. Hauptziel bei der Entwicklung war die Schaffung einer kleinen, einfachen, verständlichen und schnellen Simulationsumgebung, die auf dem Konzept eines “Flows” als kleinster zu simulierender Einheit beruht und sich hierin von den verbreiteten paketbasierten Simulationstools wie beispielsweise ns-2 [ns-2] absetzt. Hierdurch werden einmal die in unserem Zusammenhang nicht weiter interessanten Details der tieferen Schichten verborgen, und zugleich wird der Simulator sehr viel schneller als im paketbasierten Fall.

Auf die genauen Details der Umgebung soll hier nicht weiter eingegangen werden, sie sind in [Schw99] und [Sp99] ausführlich beschrieben. Eine wichtige Eigenheit ist der verwendete “Design-by-Interfaces”-Ansatz, wonach Interfaces und Implementation getrennt designt wurden. Die Implementierung wird also nur über die entsprechenden Interfaces genutzt und kann daher unabhängig von diesen leicht geändert werden, falls dies notwendig wird. Als Routing-Verfahren für die Wahl des Pfades von Sender zu Empfänger wurde schließlich das bekannte Distance-Vector Routing verwendet, das im wesentlichen in jedem Knoten eine Tabelle verwaltet, in der die kürzeste bekannte Distanz zu jeder beliebigen Destination sowie der hierfür zu verwendende Pfad aufgeführt ist [Tan96].

### 8.6.2 Festlegung des Simulationsszenarios

Die den Simulationen zugrundeliegende Topologie wurde unter Berücksichtigung folgender Punkte gewählt:

- Sender, Empfänger und ISPs werden alle in Form von Netzknoten modelliert. Verschiedene Sender schicken dabei Flows an unterschiedliche Empfänger.
- Es kann zu Bottlenecks im Netz kommen.
- Ein Sender-Empfänger-Paar kann über unterschiedliche Routen erreicht werden.
- Routen können unterschiedlich lange sein, d.h. über mehr oder weniger Hops verlaufen.

Die verwendete Topologie ist in Abbildung 8-3 dargestellt. Hierbei sind die Nutzer in den Knoten 0-3, 5, 9 und 11-13 plaziert, die übrigen Knoten entsprechen ISPs, die für die Benutzung ihrer Ressourcen Gebühren verlangen und diese durch Auktionen ermitteln. Die Netzbandbreiten liegen in der Regel bei 155 Mbps, können aber im sog. “Überlastfall”, d.h. bei Auftreten eines Bottlenecks, deutlich (bis hinunter zu einer Größenordnung von 2 Mbps) reduziert sein.

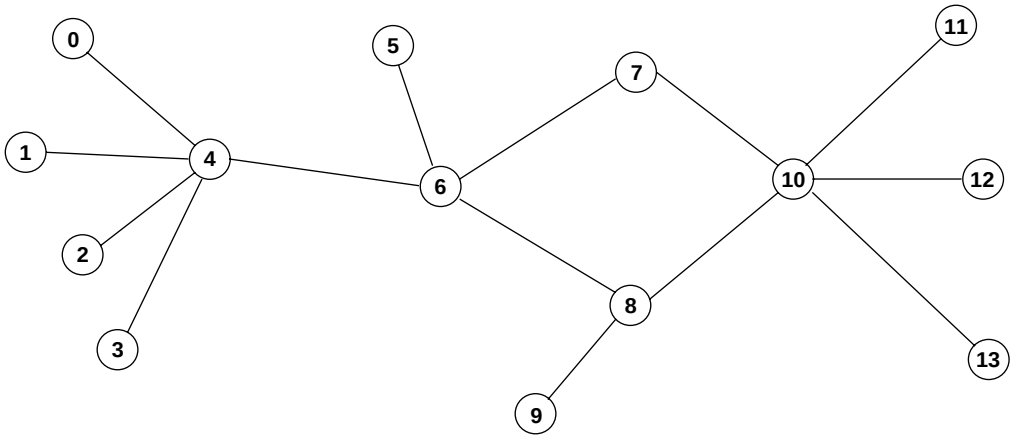


**Abbildung 8-3:** Topologie des Simulationsszenarios

Flows zwischen Sendern und Empfängern werden gemäß einer Exponentialverteilung zufällig gestartet, und zwar im Mittel alle 4 Sekunden einer; ihre Dauer ist ebenfalls exponentialverteilt mit Mittelwert 600 Sekunden. Aus Gründen der Übersichtlichkeit ist die Bandbreitenanforderung bei jedem Flow auf 256 Kbps festgelegt. Die Simulationsdauer umfaßt jeweils ca. 30 Minuten. Zusätzlich wurde jeweils auch noch der Signalisierungsverkehr in die Simulation miteinbezogen. Die Signalisierung erfolgte dabei im Anschluß an das RSVP-Protokoll mit den in Abschnitt 8.5.4 eingeführten Erweiterungen.

## 8.6.3 Simulationsergebnisse

Aufbauend auf dem vorgegebenen Szenario wurde eine erste simulative Untersuchung und Bewertung der Auktionsmechanismen durchgeführt. Im folgenden werden einige der Ergebnisse exemplarisch herausgestellt, eine detailliertere Beschreibung findet sich in [Sp99]. Wir beschränken wir uns hierbei jeweils auf den Überlastfall, der auf eine Reduzierung der Bandbreite im Netz des ISPs von Knoten 4 zurückzuführen ist.

### Marktpreis und Spending-Caps im Überlastfall

In einem ersten Schritt wurde mit Hilfe des entwickelten Simulationstools die Entwicklung des Marktpreises in Abhängigkeit vom zugrundeliegenden Spending-Cap untersucht, wenn der ISP von Knoten 4 ein Bottleneck im Netz repräsentiert. Hierbei lag das Spending-Cap der einzelnen Nutzer entweder zwischen 1000 und 1010, oder zwischen 1000 und 1050, oder schließlich zwischen 1000 und 1100, und wurde für jeden einzelnen Flow zufällig aus diesen Intervallen gewählt.

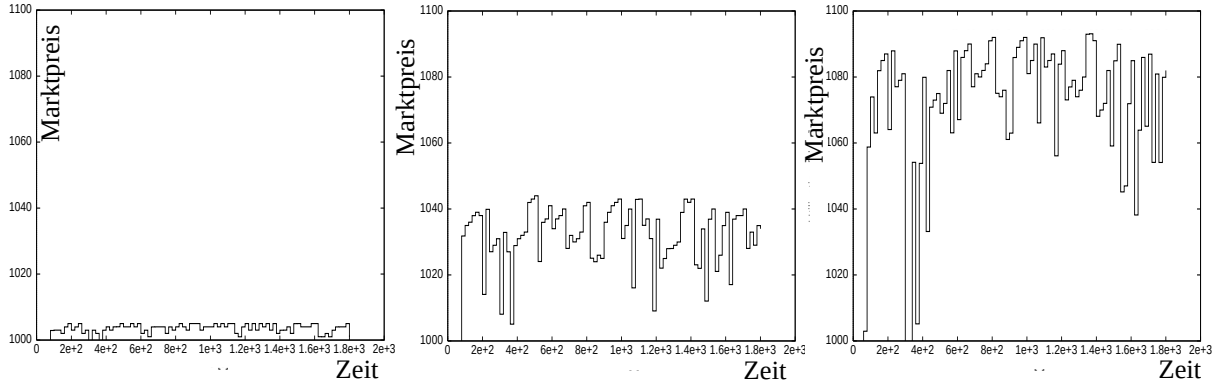


**Abbildung 8-4:** Marktpreisentwicklung im Fall einer Überlastung von Knoten 4. Die Spending-Caps liegen dabei im Intervall [1000,1010] (links) bzw. [1000, 1050] (Mitte) bzw. [1000,1100] (rechts).

Abbildung 8-4 zeigt in diesen drei Fällen die Entwicklung des Marktpreises für den überlasteten ISP. Wir sehen deutlich, daß der Marktpreis in allen Fällen das gesamte verfügbare Spektrum durchquert, insbesondere auch im dritten Fall nicht ständig an der oberen Grenze verbleibt.

### Klassische Auktion vs. CHiPS

Für einen Vergleich zwischen einer klassischen Second-Price-Auktion und CHiPS wird exemplarisch der Fall von zwei Nutzern (Knoten 0 und Knoten 1) betrachtet, wobei Nutzer 0 ein höheres Budget zur Verfügung hat als Nutzer 1. Wir sehen in Abbildung 8-5, daß dies wie erwartet die Verdrängung von Nutzer 1 zur Folge hat.

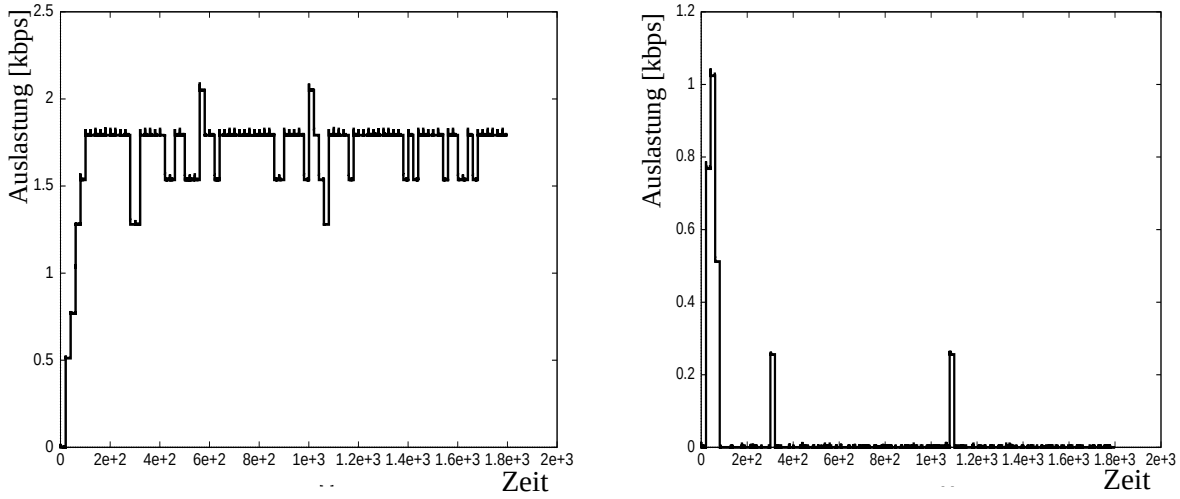


**Abbildung 8-5:** Auslastung der Teilverbindung zwischen Knoten 0 und 4 (links) bzw. Knoten 1 und 4 (rechts). Nutzer 0 hat das doppelte Budget von Nutzer 1 zur Verfügung.

In Zahlen ausgedrückt ergeben sich die Werte von Tabelle 8-6. Wir sehen, daß ohne CHiPS etwa 60% der akzeptierten Flows vorzeitig abgebrochen werden müssen, so daß de facto noch 26 regulär zu Ende gebrachte Verbindungen übrig bleiben. Diese Zahl erhöht sich bei Verwendung von CHiPS deutlich. Zudem zeigt ein Vergleich der von Nutzer 0 bzw. Nutzer 1 insgesamt entrichteten Gebühren, daß CHiPS auch eine Verbesserung hinsichtlich der Fairneß zwischen beiden Nutzern darstellt.

| Link 4–6 | ohne CHiPS | mit CHiPS |
|---|---|---|
| Akzeptierte Gespräche | 66 | 34 |
| Verlorene Gespräche | 392 | 424 |
| Unterbrochene Gespräche | 40 | 1 |
| Einkommen Knoten 0 | 2.8870 E+11 | 2.9446 E+11 |
| Kosten Nutzer 0 | 2.8837 E+11 | 2.4919 E+11 |
| Kosten Nutzer 1 | 0.0009967 E+11 | 0.4588 E+11 |

**Tabelle 8-6:** Vergleich Überlastfall mit und ohne CHiPS

### Zur Incentive Compatibility von CHiPS

Zuletzt wird noch untersucht, inwiefern der CHiPS-Ansatz die Incentive Compatibility der unterliegenden Second-Price-Auktion erhält. Konkreter formuliert handelt es sich (vgl. Abschnitt 8.2.2) um die Frage, ob ein Nutzer aus der Abgabe eines Gebots, das sich von dem Wert $\vartheta_i$, den die Ressource für ihn tatsächlich hat, unterscheidet, einen Vorteil ziehen kann, insbesondere ob sich dadurch der Wert der Utility-Funktion gem. (8.4) erhöht.

Hierzu wurde eine neue Variable definiert, der sogenannte "*Strategie-Faktor*" $\sigma_i$. Nutzer *i* ermittelt demzufolge sein Gebot $\gamma_i$ aus dem Produkt von tatsächlichem Wert $\vartheta_i$ und Strategie-Faktor $\sigma_i$:

$$\gamma_i = \sigma_i \cdot \vartheta_i \tag{8.9}$$

Abbildung 8-7 zeigt den Wert der Nutzer-Utility in Abhängigkeit vom Strategie-Faktor $\sigma_i$. Das Maximum genau beim Wert $\sigma_i = 1$ läßt sich dahingehend interpretieren, daß der Wert der Utility-Funktion (8.4) maximal wird, wenn der Nutzer genau den tatsächlichen Wert der Ressource als Gebot abgibt. Dies bedeutet, daß sich die wichtige Eigenschaft der Incentive Compatibility, wie sie die zugrundeliegenden Second-Price-Auktionen besitzen, auf den Mechanismus von CHiPS übergeht. Damit ist eine wichtige Voraussetzung dafür geschaffen, mit dem neuentwickelten Auktionsschema auch im Fall von Multiproviderszenarien zu optimalen oder Pareto-optimalen Lösungen der Ressourcenverteilung zu kommen.

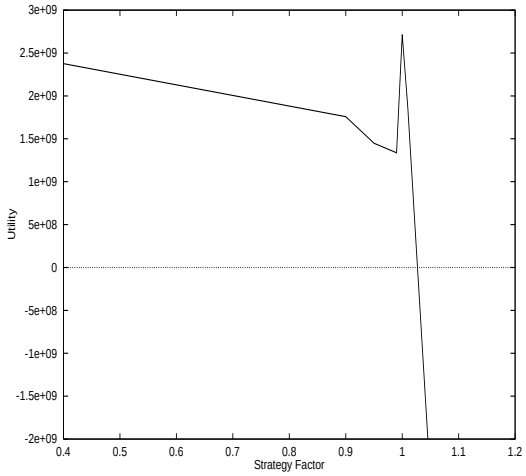


**Abbildung 8-7:** Erhaltung der Incentive Compatibility unter CHiPS

## 8.7 Fazit

Das vorliegende Kapitel beschäftigte sich mit der Anpassung geeigneter Auktionsschemata an die besonderen Bedingungen, die in Multiprovider-Szenarien vorliegen. Zur Lösung der hier auftretenden Probleme wurde ein neuer Auktionsmechanismus vorgeschlagen, der sich als vielversprechender Ansatz für eine auktionsbasierte Preismodellierung von integriertem Internetverkehr entpuppt hat. Damit stellt er nach dem Konzept der Preisfunktionen den zweiten denkbaren Kandidaten für den Einbau in eine reale Charging-und-Accounting-Plattform dar. Bis es jedoch soweit ist, sind noch einige weitere Anpassungen vorzunehmen, auf die im folgenden Kapitel noch einmal ein besonderes Licht geworfen wird.

KAPITEL 9

# *Ergänzende Bemerkungen zur Modellierung von Internet-Tarifen*

## 9.1 Zur Benutzerakzeptanz dynamischer Preismodelle

Einer der wichtigsten Gesichtspunkte beim Entwurf jedes dynamischen Tarifschemas, das für die praktische Verwendung in einem Kommunikationssystem wie dem Internet geeignet sein soll, betrifft die Frage, ob der letztendlich davon betroffene Nutzer überhaupt willens bzw. in der Lage ist, mit Tarifen umzugehen, die unvorhergesehenen Schwankungen unterliegen können, und in welcher Weise man eventuell solche Schwankungen benutzerfreundlich abfedern könnte. Zur Klärung der ersten Frage wird in Abschnitt 9.1.1 kurz auf ein entsprechendes Experiment der Universität von Berkeley hingewiesen, während der folgende Abschnitt einen einfachen Ansatz vorschlägt, der es dem Nutzer in die Hand gibt, das Ausmaß möglicher Preisschwankungen seinen Bedürfnissen und Vorstellungen anzupassen.

### 9.1.1 Das INDEX-Experiment

Unter dem Namen INDEX (für Internet Demand Experiment) läuft an der Universität von Berkeley (Kalifornien) seit April 1998 ein Markt- und Technologieexperiment, das herausfinden soll, inwiefern es Internet-Nutzer zu schätzen wissen, wenn man ihnen eine Auswahl unter qualitativ unterschiedlichen Optionen für den Internetzugang zu entsprechend unterschiedlichen Preisen anbietet. Die derzeit etwa 70 Teilnehmer haben dabei zunächst einen ausführlichen allgemeinen Fragebogen auszufüllen, der für die spätere statistische Auswertung der Resultate von Nutzen ist. In der Folge nehmen sie an einer Reihe von Experimenten von jeweils sechs bis zehn Wochen Dauer teil, während derer ihnen ein bestimmtes “Menü” von Kombinationen aus Zugangsqualität und Preisen für den Internetzugang angeboten wird, wobei in jedem Experiment ein neues Preismodell ausprobiert wird. Bislang in INDEX verwendete Preismodelle basierten u.a. auf Volumen, Volumen plus Zugangskapazität, symmetrischer oder asymmetrischer Bandbreite, auch wurden sog. “self-selecting tariffs” (eine konvexe

Kombination von Verbindungszeit und Volumen, die wöchentlich geändert werden konnte) und schließlich eine verbesserte Version des “flat rate”-Schemas (unbeschränkter Zugang mit niedriger Kapazität plus Aufschlag für höhere Zugangskapazität) implementiert. Alle Preismodelle werden dabei explizit im Gegensatz zu dem bei heutigen Internet-Providern weitverbreiteten einfachen “flat rate”-Schema entworfen.

Für detaillierte Ergebnisse dieser Experimente wird auf [Ed99], [EV99] und auf die INDEX-Homepage[1] verwiesen. Allgemein läßt sich jedoch feststellen, daß die Nutzer sehr wohl zwischen den unterschiedlichen verfügbaren Dienstqualitäten differenzieren und eine Auswahl treffen, die ihren Bedürfnissen entspricht. Die INDEX-Daten zeigen, daß die Nachfrage sehr sensitiv im Hinblick auf Preise und angebotene Qualität reagiert, daß innerhalb der Nutzer unerwartet große Unterschiede in der Nachfragestruktur bestehen und daß auch das angebotene Preismodell (z.B. volumen- oder zeitbasiert) für die Nachfrage eine große Rolle spielt. Darüberhinaus stellt sich heraus, daß sich das heute unter den ISPs dominierende “flat rate”-Schema durch Ressourcenverschwendung, Unfairneß unter den Nutzern und Einnahmeverluste für die ISPs auszeichnet.

Als konstruktiver Gegenvorschlag wird innerhalb von INDEX deshalb ein “Alternativ-ISP” entwickelt, der auf differenziertem Dienstangebot zu nutzungsbasierten Preisen, die ihrerseits die Kosten für die Ressourcen reflektieren, beruht. Die Zugangsgebühren sollen dabei deutlich unter den entsprechenden “flat rate”-Gebühren liegen, obwohl die Qualität letztendlich höher ist. Zudem erlauben entsprechende Feedback-Mechanismen den Nutzern eine bessere Kontrolle über ihr Konsumverhalten. Letztendlich hat der bisherige Verlauf des INDEX-Projektes also gezeigt, daß sowohl die Technologie wie auch der Markt für die Verwendung dynamischer Tarife für die Internet-Nutzung vorhanden ist.

### 9.1.2 Ein Parameter zur Begrenzung von Preisschwankungen

Auch wenn die geschilderten Ergebnisse des INDEX-Projektes die grundsätzliche Offenheit der Nutzer hinsichtlich variabler Tarifschemata erkennen lassen, so ist damit noch keine Aussage über die entsprechende zeitliche Auflösung, also z.B. den Umfang plötzlich eintretender Preisschwankungen, getroffen. Diesbezügliche Diskussionen lassen immer wieder den Wunsch nach einem zusätzlichen Parameter erkennen, über den der Nutzer spezifizieren kann, in welchem Umfang er bereit ist, seine Gebühren dem momentanen Marktgeschehen anzupassen. Als einfachen Ansatz schlagen wir daher einen Parameter $\alpha \in [0, 1]$ vor, der vom Nutzer vorab einzustellen ist. Dabei entspricht der Wert $\alpha = 0$ einem Preis, der über die gesamte Zeit der Verbindung stets konstant bleibt, während $\alpha = 1$ einem Tarif entspricht, der ständig (d.h. unter Verwendung von RSVP beispielsweise ca. alle 30 Sekunden, also dem Auktionsintervall bzw. Auffrischungsintervall von RSVP) der Marktentwicklung folgt.

1. http://www.INDEX.Berkeley.edu

Implementieren ließe sich dieser Parameter am einfachsten in Form des “exponential averaging”-Ansatzes[2]. Demnach ergibt sich nach Ablauf der Auktionsperiode *t* der neue dem Benutzer *n* in Rechnung zu stellende Preis $p_t$ aus dem Preis der vorangegangenen Periode $p_{t-1}$ und dem momentanen Marktpreis $m_t$ zu

$$p_t = \alpha_n \cdot m_t + (1 - \alpha_n) \cdot p_{t-1}, \tag{9.1}$$

wobei $\alpha_n$ durch den Benutzer festgelegt wird. Initialisiert der ISP diese Formel, indem er für $p_0$ einen geeigneten Durchschnittspreis wählt (z.B. in Abhängigkeit von der Tageszeit, oder aber wirklich einfach die durchschnittlich für eine Verbindung des betroffenen Typs erhobene Gebühr), dann bleibt bei der Preis bei der Wahl von $\alpha_n = 0$ stets gleich diesem Durchschnittspreis, während er sich für steigendes $\alpha$ immer enger an der aktuellen Marktentwicklung orientiert. Der Einbau dieses Parameters in die CATI-Plattform ist vorgesehen (vgl. Abschnitt 9.3).

## 9.2 Auf dem Weg zum Edge Pricing - oder: “Effektive Bandbreite” einmal anders

Ein weiterer wichtiger Aspekt bei der Entwicklung eines effizienten Preismodells für den Einsatz in einer realen Umgebung betrifft die Frage, an welchen Stellen im Netz eine Preisermittlung stattfinden soll. Der bislang favorisierte Ansatz, den Preis für jeden von einer Multiprovider-Verbindung betroffenen ISP bzw. sogar für jede Teilverbindung lokal und unabhängig voneinander zu ermitteln, kann sehr schnell zu Skalierungsproblemen führen. Das in Abschnitt 6.2.1 erläuterte Edge Pricing-Paradigma löst dieses Problem, indem es vorschlägt, die Gebühr für eine Verbindung nur einmalig, und zwar am Beginn der Verbindung (d.h. beim sog. Access-ISP) zu ermitteln, ohne jedoch konkret zu werden, wie das denn vor sich gehen könnte.

Die Ausarbeitung dieses Ansatzes ist Gegenstand aktueller und zukünftiger Forschungsarbeiten. An dieser Stelle soll nur kurz skizziert werden, auf welchem Weg man mit Hilfe der bisher entwickelten Werkzeuge zu einer befriedigenden Lösung kommen könnte. Hierzu gehen wir wiederum davon aus, daß Second Price-Auktionen grundsätzlich ein wirksames Instrument zur Bestimmung der aktuellen Marktsituation für eine einzelne Teilverbindung darstellen. Wie in Kapitel 8 ausführlich beschrieben, hat ein Nutzer, der eine Multiprovider-Verbindung aufbauen und unterhalten möchte, ein gewisses Globalbudget zur Verfügung, das er für die einzelnen lokalen Versteigerungen geeignet aufteilen muß. Ist dieses Globalbudget zu niedrig angesetzt oder teilt er es falsch auf die Einzelgebote auf, dann wird er mit seinem Verbindungswunsch keinen Erfolg haben.

2. Diese Idee wird beispielsweise beim CPU-Scheduling zur Schätzung erwarteter Prozeßzeiten verwendet, vgl. [SPRS96]

Formal entspricht diese Aufteilung also einer Abbildung

$$\Gamma: \begin{cases} \mathbb{R} \rightarrow \mathbb{R}^N \\ B \rightarrow \Gamma(B) = (\gamma_1, \gamma_2, \ldots, \gamma_N) \end{cases} \tag{9.2}$$

mit

$$\gamma_n = \gamma_n(d_n, M_n) \qquad \forall\ n = 1, 2, \ldots, N \tag{9.3}$$

des Budgets $B$ auf die Gebote $\gamma_n$ für die $N$ einzelnen lokalen Auktionen. Jedes Gebot betrifft dabei eine Auktion um eine Bandbreite $d_n$ bei Vorliegen einer aktuellen Marktsituation, die sich durch einen Parameter $M_n$ charakterisieren läßt (z.B. den aktuellen Marktpreis per Ressourcen-Einheit).

Edge Pricing ist in gewissem Sinn die Umkehrung der Abbildung (9.2). Anstatt $N$ einzelne Gebühren für die $N$ Teilverbindungen zu ermitteln und zu bezahlen, soll der Access-ISP einmalig eine entsprechende Gebühr ermitteln und einziehen. Bleibt man auf dem Fundament auktionsbasierter Preismodelle, so entspricht dies einer beim Access-ISP durchgeführten Auktion, die nunmehr die gesamte Verbindung betrifft.

Um welches Gut dreht sich nun diese Auktion? Eine mögliche Antwort auf diese Frage besteht in einer "effektiven Bandbreite"[3] $\Delta = \Delta(d_1, \ldots, d_N; M_1, \ldots, M_N)$, die von allen Bandbreiten der Teilverbindungen und den entsprechenden Marktsituationen abhängt. Jeder an einem bestimmten Access-ISP angeschlossene Nutzer $i$ nimmt dann an der Auktion dieses Access-ISPs teil und bietet für das seinem Verbindungswunsch entsprechende $\Delta_i$. Insgesamt wird bei dieser Auktion eine entsprechend zu bestimmende Gesamtbandbreite $\Delta_{ges}$ versteigert. Beide Ansätze - verteilte Auktionen und Edge Pricing - sind genau dann äquivalent, wenn genau die Nutzer, denen aufgrund eines genügend hohen Budgets der Verbindungsaufbau über eine verteilte Auktion gelingt, auch bei der Edge Pricing-Auktion Erfolg haben, und umgekehrt.

Die genaue Form dieser Abbildungen $\Delta_i$ bzw. $\Delta_{ges}$ ist freilich noch offen. Immerhin besteht die Hoffnung, durch Simulation geeigneter Szenarien mittels der erweiterten FlowSim-Umgebung Anhaltspunkte bzw. Randbedingungen für das Aussehen der Abbildungen zu erhalten (z.B. entspricht der Fall eines nicht ausgelasteten Netzes $\Delta \approx 0$). Weitere Ergebnisse in dieser Richtung müssen allerdings künftiger Forschungsarbeit vorbehalten bleiben.

3. Der Begriff "effektive Bandbreite" wird hier bewußt anders verwendet als im üblichen ATM-Kontext (vgl. z.B. [Kel91a]). Die dahinterstehende suggestive Idee ist jedoch dieselbe, nämlich Reduktion einer vieldimensionalen Information auf einen charakteristischen Parameter, der die lineare Behandlung des Problems ermöglicht. Daher verzeihe man die Anleihe bei der ATM-Welt.

## 9.3 Charging-und-Accounting-Technologie im Internet: Das Projekt CATI

Nachdem in den vorhergehenden Abschnitten ein erster Ausblick auf zukünftige Untersuchungen gegeben wurde, dienen die folgenden Bemerkungen der Einbettung der einzelnen Preismodelle in den größeren Zusammenhang eines derzeit laufenden Forschungsprojektes, das - wie schon mehrmals erwähnt - das Design und die Realisierung eines funktionsfähigen Charging-und-Accounting-Tools zum Ziel hat. Nach einem Überblick über das Gesamtprojekt wird dabei noch kurz auf das weitere Schicksal der entwickelten Preismodelle im Rahmen dieses Projektes eingegangen.

Über das seit Jahren zu beobachtende exponentielle Wachstum und den vieldiskutierten Übergang zu IPv6 [Hui96] hinaus ist in letzter Zeit im Internet eine Reihe von weiteren Veränderungen zu beobachten. Zum einen nimmt der bislang übliche hohe Anteil staatlicher Finanzierung mehr und mehr ab, um einer zunehmenden Kommerzialisierung des Internet Platz zu machen. Parallel dazu entwickelt sich das Internet mehr und mehr vom “Best Effort”-Netz hin zum Angebot einer Reihe von qualitativ höherstehenden Diensten, um eine breitere Palette von Anwendungen (wie etwa IP-Telefonie, Video-Conferencing, interaktives Fernsehen, Virtual Private Networks VPN und dergleichen mehr) unterstützen zu können und dadurch eine weitere Differenzierung des Kommunikationsmarktes zu ermöglichen.

Vor diesem Hintergrund ergibt sich ein schnell steigender Bedarf nach einer geeigneten Charging-und-Accounting-Infrastruktur für das Internet [SFPN98]. Während die “*Accounting*”-Komponente dabei, wie bereits kurz erwähnt, für das Sammeln von Informationen über Nutzung und Konsum von Netzressourcen verantwortlich ist, versteht man unter “*Charging*” allgemein die Transformation dieser gesammelten Parameter in monetäre Einheiten, z.B. mit Hilfe geeigneter Preismodelle. Ziel des vom Schweizer Nationalfond (SNF) Bern geförderten Projektes “CATI” (Charging and Accounting Technology for the Internet) ist das Design, die Implementation und Evaluation einer derartigen Plattform, die exemplarisch anhand eines IP-Telefondienstes nachweisen soll, daß Gebührenerhebung für die Internet-Nutzung sowohl technologisch machbar als auch ökonomisch sinnvoll ist.

In Kürze lassen sich die Hauptziele des Projektes folgendermaßen zusammenfassen (vgl. [SBGP99]):

- Design und Implementation von Charging- und Accounting-Mechanismen auf der Grundlage heute verfügbarer sicherer und reservationsbasierter Internet-Protokolle;
- Design und Implementierung eines VPN-Konfigurationsdienstes einschließlich Charging- und Accounting-Funktionalitäten;
- Entwicklung allgemeiner Support-Funktionalitäten in Form von Application Programming Interfaces (API) für internetbasierten E-Commerce;

- Entwicklung eines IP-Telefons als Demonstrator, das auf die implementierten Charging- und Accounting-Funktionalitäten sowie auf die QoS-Unterstützung in Internet-Protokollen zurückgreift [SFJ+99];
- Spezifikation und Untersuchung von Business-Modellen für Internet-Dienste einschließlich Kosten- und Tarifschemata für Best-Effort- wie IntServ-/DiffServ-Architekturen;
- Bewertung der entwickelten Charging- und Accounting-Mechanismen, Preismodelle und Tarifschemata für die Demonstrator-Anwendung wie auch für regulären Netzverkehr.

Die in den vorhergehenden Kapiteln erarbeiteten Preismodelle sind eng mit diesem Projekthintergrund verknüpft. Dabei hat sich herausgestellt, daß sowohl der Ansatz über die Preisfunktionen wie auch die Weiterentwicklung des Auktionsverfahrens für eine Verwendung in der entstehenden Plattform brauchbar ist. Hinsichtlich der RSVP-Erweiterung sind, wie bereits am entsprechenden Ort dargestellt, für beide Ansätze geeignete neue RSVP-Objekte definiert und implementiert worden. Was die Integration in die Gesamtarchitektur betrifft, so läßt sie sich wohl einfacher für die Preisfunktionen durchführen, da hierbei lediglich pro ISP bzw. Teilverbindung die momentane Auslastung festzustellen ist, woraufhin eine einfache Datenbankabfrage bereits den gesuchten lokalen Preis ergibt. Demgegenüber scheint der Implementierungsaufwand für eine CHiPS-Auktion ungleich größer, da neben den entsprechenden Auktionatoren auch noch jeder Nutzer davon betroffen ist. Natürlich erhält man dadurch letztendlich ein Preismodell, das um vieles flexibler und genauer auf den sich ändernden Markt reagiert als das über wiederholte Approximationen gewonnene Alternativmodell der Preisfunktionen. Eine Entscheidung zwischen diesen beiden Ansätzen muß daher nicht zuletzt darauf beruhen, wie sich das Verhältnis aus Dynamik, Präzision und Aufwand bei der Preisbestimmung aus der Sicht des Nutzers wie auch des ISPs, also letztlich der Abnehmer und Kunden des entstehenden Systems, darstellt.

KAPITEL 10

# Schlußbemerkung

Fassen wir zusammen: Die vorliegende Arbeit ist Untersuchungen gewidmet, die sich an der Schnittstelle zwischen mathematisch-theoretischer Modellbildung und dem Einsatz derartiger Modelle in real implementierten Plattformen und Tools ansiedeln lassen. Insbesondere geht es dabei um die Frage einer geeigneten *dynamischen Modellierung* von Phänomenen, die trotz offenkundiger Zeitabhängigkeit bislang mehr oder weniger rein statisch modelliert wurden. Vor diesem Hintergrund betrachten wir zwei Beispiele für derartige Fragestellungen, die dem Bereich der *Verkehrs- bzw. Preismodellierung in Kommunikationssystemen* entnommen sind.

Das erste Beispiel betrifft die dynamische Modellierung von empirisch ermitteltem Verkehrsaufkommen, wie es in Kapitel 2 anhand von Location Updates in einem GSM-Netz exemplarisch vorgestellt wird. Kapitel 3 beschreibt die (leider fehlgeschlagenen) Versuche, dieses Referenzproblem mit klassischen autoregressiven Modellen anzugehen, bevor sich in Kapitel 4 der Blick auf die sogenannte TES-Methode richtet. Nach der Einführung dieses Modells in der üblichen Form wird eine Erweiterung vorgeschlagen, die spezifisch auf die Modellierung von *periodischem* Verkehr unter möglichst geringem Aufwand ausgerichtet ist. Sodann wird nachgewiesen, wie sich damit wichtige Charakteristika einer empirischen Messung (Autokorrelationsstruktur, Randverteilung und insbesondere sogar der zeitliche Verlauf) so gut nachbilden lassen, daß der Einsatz dieser Erweiterung in einer echtzeitfähigen GSM-Testumgebung möglich wird. Kapitel 5 zeigt zum Abschluß dieses Teils einige Richtungen für weiterführende Untersuchungen auf. Insgesamt bleibt als Resultat festzuhalten, daß nunmehr ein *neuartiges Verfahren zur Modellierung von periodischem Verkehr* zur Verfügung steht, das bei extrem *niedriger Komplexität* Resultate liefert, die eine *ausgezeichnete Übereinstimmung aller praktisch relevanten statistischen Parameter von Modell und Originalmessung* aufweisen.

Hierauf wenden wir uns der Frage nach dynamischen Preismodellen im Internet zu. Der Stand der Dinge auf diesem Gebiet wird in Kapitel 6 zusammengefaßt, bevor in Kapitel 7 ein mathematisches Modell entwickelt wird, das erlaubt, einen Zusammenhang zwischen der Auslastung

einer Ressource und der für ihre Nutzung zu entrichtende Gebühr abzuleiten. Die neuartige Verwendung eines asymptotischen Approximationsansatzes verringert auch in diesem Fall die Komplexität der Lösung derart, daß es dadurch möglich wird, erstmals *Preisfunktionen für Internet-spezifische Szenarien* mit hohen Bandbreiten und unterschiedlichen Verkehrsklassen numerisch anzugeben.

Alternativ dazu geht Kapitel 8 die Problematik im Kontext von auktionsbasierten Tarifmodellen an und stellt CHiPS (Connection-Holder-is-Preferred-Scheme) als *neuen Auktionsmechanismus für Verbindungen* vor, die *über mehrere Internet-Provider* laufen. Basierend auf Simulationsergebnissen wird gezeigt, daß dieser Ansatz in der Lage ist, die mit dem Einsatz von Auktionen in Multiprovider-Szenarien neu auftretenden Probleme befriedigend zu lösen. In Kapitel 9 werden die Einbindung beider Ansätze in laufende Projekte verdeutlicht und offengebliebene Fragen angerissen. Als ein zentrales Ergebnis dieses zweiten Teils der Arbeit bleibt festzuhalten, daß es gelungen ist, ein mathematisches Verfahren zur Approximation von auslastungsabhängigen Preisfunktionen für Einzelressourcen anzugeben, dessen Komplexität von der Kapazität der Ressource unabhängig ist. Ferner wurde gezeigt, auf welche Weise sich ein für die Preisbestimmung von Einzelressourcen bekanntermaßen geeigneter Auktionsmechanismus auf die im Internet herrschenden besonderen Verhältnisse anpassen läßt, ohne seine guten Eigenschaften zu verlieren.

Daß und wie auf jedem der behandelten Gebiete noch eine Fülle offener Fragen die weitere Arbeit daran stimuliert, wurde bereits an den entsprechenden Stellen deutlich gemacht. Die derzeitige Resonanz auf die vorgeschlagenen Ansätze zeigt sich am deutlichsten an ihrem vorgesehenen Einsatz im Rahmen verschiedener laufender Projekte unter Industriebeteiligung. Was darüber hinaus in jedem Fall als eindringliche Erfahrung bleibt, ist das Erlebnis einer gegenseitigen Befruchtung von Theorie und Praxis, das als solches Ansporn und Anspruch zugleich geworden ist.

ANHANG A

# *Mathematische Grundlagen und Ergänzungen*

## A.1 Zum mathematischen Hintergrund von TES

**Lemma A.1:** *Iterierte Gleichverteilung*

Sei $U$ eine (0,1)-gleichverteilte Zufallsvariable (d.h. $U \sim U(0, 1)$) und $V$ eine davon unabhängige Zufallsvariable mit beliebiger Verteilung. Dann ist $W = \langle U + V \rangle_{\text{mod } 1}$ ebenfalls (0,1)-gleichverteilt.

***Beweis:*** Zunächst wird gezeigt, daß gilt: $X = \langle U + c \rangle \sim U(0, 1)$ für jedes reelle $c$. Hierzu sei o.B.d.A. $c \in [0, 1)$, denn für jedes feste reelle $c$ gilt: $\langle U + c \rangle = \langle U + \langle c \rangle \rangle$. Für $x \in [0, 1]$ gilt:

$$P\{X \le x\} = P\{U + c < 1, U + c \le x\} + P\{U + c \ge 1, U + c - 1 \le x\} = A + B$$

mit

$$A = P\{U + c < 1, U + c \le x\} = (x - c) \cdot 1_{[c, 1)}(x) \text{ und}$$
$$B = P\{U + c \ge 1, U + c - 1 \le x\} = x \cdot 1_{[0, c)}(x) + c \cdot 1_{[c, 1)}(x).$$

Bildet man nun die Summe aus $A$ und $B$, so erhält man

$$P\{X \le x\} = A + B = x$$

und somit $X \sim U(0, 1)$.

Sei nun $x, y \in [0, 1]$. Dann gilt:

$$P\{X \le x | V = y\} = P\{\langle U + y \rangle \le x\} = x,$$

unabhängig von $y$. □

**Definition A.2:** *Pseudo-Inverse*

Es sei $F: \mathbb{R} \to \mathbb{R}$ eine schwach monoton wachsende, rechtsseitig stetige Funktion und $I(F) = inf\{F(x) | x \in \mathbb{R}\}$, $S(F) = sup\{F(x) | x \in \mathbb{R}\}$. Dann ist auf dem offenen Intervall $(I(F), S(F))$ die Pseudo-Inverse $F^{-1}$ von $F$ definiert durch

$$F^{-1}(y) = inf\{x \in \mathbb{R} | F(x) \geq y\},\ I(F) \leq y \leq S(F).$$

***Anmerkung:*** Der grundlegende Unterschied zwischen der "normalen" Inversen und der Pseudo-Inversen ist die Handhabung von Sprungstellen. Die Inverse ist nur für Funktionen ohne Sprungstellen definiert. Bei der Pseudo-Inversen werden Sprungstellen zu Konstanten und Konstanten zu Sprungstellen. Außerdem ändert sich die Stetigkeit. Eine linksseitig stetige Funktion wird durch die Pseudo-Inverse zu einer rechtsseitig stetigen und umgekehrt.

Die Pseudo-Inverse besitzt einige grundlegende Eigenschaften, die durch Konvergenzbetrachtungen leicht bewiesen werden können (vgl. [Moh99]), insbesondere

$$F(F^{-1}(y)) \geq y \qquad \forall I(F) \leq y \leq S(F) \tag{A.1}$$

$$F^{-1}(F(x)) \leq x \qquad \forall I(F) \leq F(x) \leq S(F) \tag{A.2}$$

$$y \leq F(x) \Leftrightarrow F^{-1}(y) \leq x, \text{ für } I(F) < F(x),\ y < S(F) \tag{A.3}$$

**Lemma A.3:** *Inversionsmethode*

Sei $F$ eine beliebige Verteilungsfunktion und $U$ eine (0,1)-gleichverteilte Zufallsvariable. Dann besitzt die Zufallsvariable

$$X = \begin{cases} F^{-1}(U), & \text{für } U \in (0, 1) \\ 0, & \text{sonst} \end{cases} \tag{A.4}$$

die Verteilungsfunktion $F$.

Ist umgekehrt $X$ eine reellwertige Zufallsvariable mit stetiger Verteilungsfunktion $F$, so ist $F(X)$ über $[0, 1)$ gleichverteilt.

***Beweis:*** Zum Nachweis der geforderten Verteilungseigenschaft genügt es, die Verteilungsfunktion der Zufallsvariablen $X$ zu betrachten. Mit (A.3) ist nun für $0 < F(x) < 1$

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

mit $P(U \notin (0, 1)) = 0$. Dies war zu zeigen.

Zum Beweis der zweiten Aussage des Lemmas kann o.B.d.A. angenommen werden, daß die Zufallsvariable $X$ bereits die durch (A.4) vorgegebene Form besitzt. Wegen der Stetigkeit von

F gilt dann nach (A.1) $F(X) = F(F^{-1}(U)) = U$, falls $0 < U < 1$. Wegen $P(0 < U < 1) = 1$ ist die Aussage bewiesen. □

Abbildung A-1 veranschaulicht die Idee der Histogramm-Inversion. $U = 0.74$ ist die zufällige Realisierung einer auf (0,1) gleichverteilten Zufallsvariablen. Da die (kumulative) Verteilungsfunktion $F$ das Einheitsintervall als Wertebereich hat, ist der Definitionsbereich der Pseudo-Inversen von $F$ ebenfalls das Intervall [0,1], wohingegen der Wertebereich von $F^{-1}$ identisch mit dem Definitionsbereich von $F$ ist, also alle möglichen Werte umfaßt, die eine nach $F$ verteilte Zufallsvariable annehmen kann. Daher ist die Zufallsvariable $X = F^{-1}(U))$ ebenfalls nach $F$ verteilt.

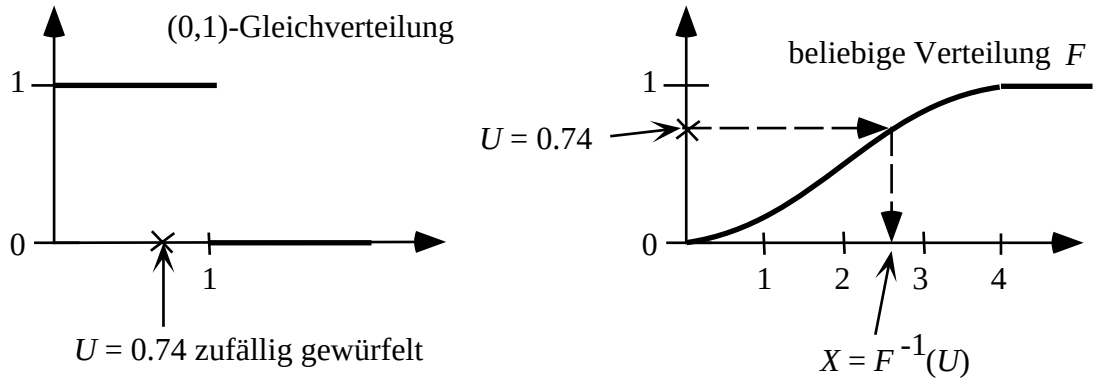


**Abbildung A-1:** Zur Illustration der Inversionsmethode

**Definition A.4:** *Randverteilung einer Stitching-Funktion*

Sei $U$ eine (0,1)-gleichverteilte Zufallsvariable auf dem Einheitsintervall und $S$ eine beliebige Stitching-Funktion. Dann heißt die Verteilung der Zufallsvariablen $V = S(U)$ *Randverteilung* der Stitching-Funktion $S$.

**Beispiel A.5:** *Berechnung der Randverteilung einer linearisierten GSF*

Als Beispiel für die Berechnung der Randverteilung einer stückweise linearen GSF gemäß Lemma 4.5 betrachten wir die GSF mit Stützstellen in den Punkten $(0, 0)$, $(0.3, 1)$, $(0.5, 0.8)$, $(0.7, 0.9)$ und $(1, 0)$ (vgl. Abbildung A-2):

$$S(x) = \begin{cases} \frac{10}{3}x, & \text{falls } x \in [0, 0.3) \\ -x + 1.3, & \text{falls } x \in [0.3, 0.5) \\ 0.5x + 0.55, & \text{falls } x \in [0.5, 0.7) \\ -3x + 3, & \text{falls } x \in [0.7\dot{}, 1). \end{cases} \tag{A.5}$$

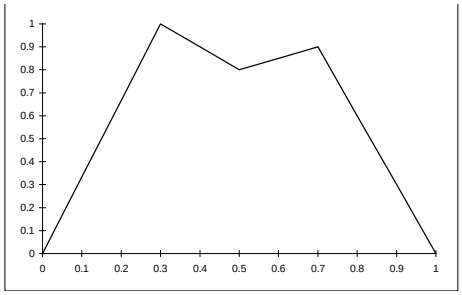


**Abbildung A-2:** Allgemeine Stitching-Funktion

Die Anwendung von Lemma 4.5 erfordert vier Iterationsschritte:

*1. Schritt:* $1 = y_1 > y_0 = 0$

$$F_1(y) = \left(\frac{y}{10/3}\right) \cdot 1_{[0,\,1]}(y) = 0.3y \cdot 1_{[0,\,1)}(y)$$

*2. Schritt:* $0.8 = y_2 < y_1 = 1$

$$F_2(y) = \left(0.5 - \frac{y - 0.3}{-1}\right) \cdot 1_{[0.8,\,1)}(y) = (y - 0.8) \cdot 1_{[0.8,\,1)}(y)$$

$$F_{neu}(y) = F_1(y) + F_2(y) = 0.3y \cdot 1_{[0,\,0.8)}(y) + (1.3y - 0.8) \cdot 1_{[0.8,\,1)}(y)$$

*3. Schritt:* $0.9 = y_3 > y_2 = 0.8$

$$F_3(y) = \left(\frac{y - 0.55}{0.5} - 0.5\right) \cdot 1_{[0.8,\,0.9)}(y) + 0.2 \cdot 1_{[0.9,\,1)}(y)$$

$$= (2y - 1.6) \cdot 1_{[0.8,\,0.9)}(y) + 0.2 \cdot 1_{[0.9,\,1)}(y)$$

$$F_{neu}(y) = F_1(y) + F_2(y) + F_3(y)$$
$$= 0.3y \cdot 1_{[0,\,0.8)}(y) + (3.3y - 2.4) \cdot 1_{[0.8,\,0.9)}(y) + (1.3y - 0.6) \cdot 1_{[0.9,\,1)}(y)$$

*4. Schritt:* $0 = y_4 < y_3 = 0.9$

$$F_4(y) = \left(1 - \frac{y - 3}{-3}\right) \cdot 1_{[0,\,0.9)}(y) + (1 - 0.7) \cdot 1_{[0.9,\,1)}(y)$$

$$= \frac{1}{3}y \cdot 1_{[0,\,0.9)}(y) + 0.3 \cdot 1_{[0.9,\,1)}(y)$$

$$F_S(y) = F_1(y) + F_2(y) + F_3(y) + F_4(y)$$
$$= 0.63y \cdot 1_{[0,\,0.8)}(y) + (3.63y - 2.4) \cdot 1_{[0.8,\,0.9)}(y) + (1.3y - 0.3) \cdot 1_{[0.9,\,1)}(y)$$

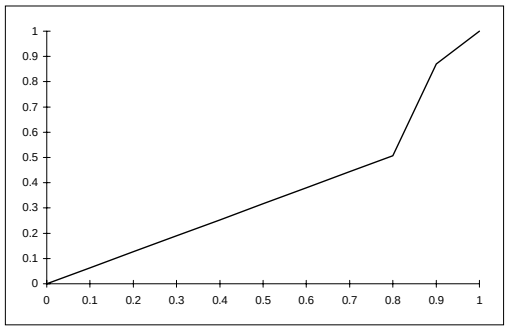


**Abbildung A-3:** Verteilungsfunktion der GSF aus Abbildung A-2

**Definition A.6:** *Gammaverteilung*

Eine Zufallsvariable $X$ heißt gammaverteilt, wenn sie für $\alpha, \lambda > 0$ eine Dichte der folgenden Form besitzt:

$$f(x) = \begin{cases} \dfrac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & \text{falls } x > 0 \\ 0, & \text{sonst} \end{cases} \tag{A.6}$$

$$\text{mit } \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx\,. \tag{A.7}$$

**Beispiel A.7:** *Eigenschaften der Gammaverteilung*

Sei $X$ gammaverteilt. Dann gilt:

$$E(X) = \frac{\alpha}{\lambda}, \tag{A.8}$$

$$Var(X) = \frac{\alpha}{\lambda^2} \tag{A.9}$$

und damit für den Variationskoeffizienten

$$c_X^2 = \frac{1}{\alpha}. \tag{A.10}$$

## A.2 Analytische Berechnung der Autokorrelationsfunktion eines einfachen TES-Modells

In (4.3) wurde bereits eine analytische Form der Autokorrelationsfunktion des TES-Modells angegeben, wie sie in Lemma A.8 nochmals konstatiert wird. Unabhängig davon werden wir anschließend zur Veranschaulichung diese analytische Form in einem einfachen Spezialfall Schritt für Schritt ableiten, um durch Vergleich mit (4.3) die Aussage von Lemma A.8 zu untermauern.

**Lemma A.8:** *Autokorrelationsfunktion von TES-Prozessen*
Sei $f_V$ eine Innovationsdichte, $D$ eine Verzerrungsfunktion und $\tilde{f}_V$ bzw. $\tilde{D}$ ihre Laplacetransformierten. Für einen gegebenen lag $\tau$ hat die Autokorrelationsfunktion folgende Form:

$$\rho_X(\tau) = \frac{2}{\sigma_X^2} \sum_{\nu=1}^{\infty} Re[\tilde{f}_V^{\tau}(\nu)]\left|\tilde{D}(\nu)\right|^2 \tag{A.11}$$

Betrachten wir nun den *Spezialfall eines unverzerrten TES-Modells mit einfacher Innovationsdichte*. Ausgangspunkt für die Berechnung ist die Tatsache, daß die Hintergrundsequenz (4.1)

$$Y_i = \langle Y_{i-1} + V_i \rangle_{\text{mod } 1} \tag{A.12}$$

als stationärer Markov-Prozeß mit $\tau$-Schritt-Übergangsdichte $g_\tau(y|x)$ interpretiert werden kann. Hierzu nehmen wir an, daß $f_V$ eine Innovationsdichte mit konvergentem Integral

$$\int_{-\infty}^{\infty} f_V(s)ds < \infty \tag{A.13}$$

ist. Diese ziemlich harmlose Bedingung ist für die spätere Anwendung der sogenannten Poissonschen Summenformel

$$\sum_{n=-\infty}^{\infty} f_V(x+n) = \sum_{-\infty}^{\infty} e^{2\pi i\nu x}\tilde{f}_V(2\pi i\nu) \tag{A.14}$$

notwendig, wobei

$$\tilde{f}_V(x) = \int_0^{\infty} e^{-sx} f_V(s)ds \tag{A.15}$$

die Laplacetransformierte der Innovationsdichte $f_V$ ist (zur Frage der Integralgrenzen vgl. die Ausführungen in Abschnitt 4.1.3, insbesondere Gleichung (4.4) und (4.5)).

Im Fall von $f_V(x) = 0$ für $x < 0$ läßt sich dies auch als

$$\tilde{f}_V(x) = \int_{-\infty}^{\infty} e^{-sx} f_V(s) ds \tag{A.16}$$

schreiben, deshalb kann man

$$\tilde{f}_V(2\pi i\nu) = \int_{-\infty}^{\infty} e^{-s2\pi i\nu} f_V(s) ds \tag{A.17}$$

als Fourierkoeffizienten zur Funktion $\sum_{n=-\infty}^{\infty} f_V(x+n)$ auffassen.

Mit Hilfe dieser Laplacetransformation dient die bedingte Dichte von $Y_\tau$, also der Hintergrundsequenz zum Lag $\tau$ (wobei $Y_0 = x$ vorgegeben sei), als die Übergangsdichte des beschriebenen Markov-Prozesses und kann nach [JM92a] berechnet werden zu

$$g_\tau(y|x) = \begin{cases} \sum_{\nu=-\infty}^{\infty} \tilde{f}_V^{\tau}(2\pi i\nu) e^{i2\pi\nu(y-x)}, & \text{falls } 0 \le y, x < 1 \\ 0, & \text{sonst} \end{cases} \tag{A.18}$$

wobei der Exponent $\tau$ in der Summe dem Ausdruck für die Laplacetransformierte der $\tau$-fachen Faltung von $f_V$ mit sich selbst entspricht. Der Beweis für diese Formel ist sehr trickreich und wird aus Gründen der Übersichtlichkeit hier übersprungen; er kann in [JM92a] bzw. (in geraffter Form) in [Moh99] nachgelesen werden. Für die konkrete Berechnungen genügt die Verwendung des Realteils

$$Re\ g_\tau(y|x) = \begin{cases} 1 + 2\sum_{\nu=1}^{\infty} Re[\tilde{f}_V^{\tau}(2\pi i\nu) e^{2\pi i\nu(y-x)}], & \text{falls } 0 \le y, x < 1 \\ 0, & \text{sonst} \end{cases}. \tag{A.19}$$

Nach diesen einleitenden Bemerkungen wenden wir uns nun dem einfachsten Fall eines TES-Modells zu, das eine Innovationssequenz der Form

$$V_n = L + (R - L) \cdot Z_n, \tag{A.20}$$

besitzt, wobei $Z_n$ eine auf [0,1) gleichverteilte Zufallsvariable sowie $L \le R$ sei. Diese sog. "LR-Parametrisierung" [JM92a] entspricht einer Innovationsdichte (4.6) mit $N = 1$ und $a_1 = L$ bzw. $b_1 = R$. Eine alternative Parametrisierung, die uns noch nützlich sein wird, erhält man über

$$\alpha = R - L \text{ und } \varphi = \frac{R + L}{\alpha}; \tag{A.21}$$

sie firmiert unter der Bezeichnung $(\alpha, \varphi)$-Parametrisierung [JM92a] und betont die Form der Innovationssequenz, die im wesentlichen bei jedem Schritt um den "Winkel" $\varphi$ weiterspringt, wobei das Ausmaß zufälliger Abweichung davon über $\alpha$ festgelegt wird (vgl. auch Abschnitt 4.1 Fußnote 1). Invertiert man (A.21), so ergibt sich übrigens

$$L = \frac{1}{2}\alpha(\varphi - 1) \text{ bzw. } R = \frac{1}{2}\alpha(\varphi + 1) \tag{A.22}$$

Sei also $f_{(L,R)}(x)$ die entsprechende Innovationsdichte (vgl. z.B. Abbildung 4-2 links). Ihre Laplacetransformierte ergibt sich zu

$$\tilde{f}_{(L,R)}(s) = \int_{-1}^{1} e^{-sx} f_V(x)dx = \int_{L}^{R} e^{-sx} \cdot \frac{1}{R-L}dx = \frac{1}{R-L} \cdot \left[\frac{e^{-sx}}{s}\right]_{x=L}^{x=R} = \frac{e^{-sL} - e^{-sR}}{(R-L)s}, \tag{A.23}$$

d.h.

$$[\tilde{f}_{(L,R)}(2\pi i\nu)]^\tau = \left[\frac{e^{-2\pi i\nu L} - e^{-2\pi i\nu R}}{2\pi i\nu(R-L)}\right]^\tau = \left[\frac{e^{-\pi i\nu\alpha(\varphi-1)} - e^{-\pi i\nu\alpha(\varphi+1)}}{2\pi i\nu\alpha}\right]^\tau \tag{A.24}$$

unter Verwendung von (A.22). Zusammen mit der wohlbekannten Euler-Formel für die komplexe Exponentialfunktion ([BS87] S. 508f.)

$$e^{ix} = \cos x + i\sin x \Rightarrow e^{ix} - e^{-ix} = \cos x + i\sin x - \cos(-x) - i\sin(-x) = 2i\sin x \tag{A.25}$$

ergibt dies

$$\begin{aligned}
[\tilde{f}_{(L,R)}(2\pi i\nu)]^\tau &= \frac{1}{(2\pi i\nu\alpha)^\tau}[e^{-\pi i\nu\alpha\varphi}]^\tau[e^{\pi i\nu\alpha} - e^{-\pi i\nu\alpha}]^\tau \\
&= \frac{1}{(2\pi i\nu\alpha)^\tau}e^{-\pi i\nu\alpha\varphi\cdot\tau} \cdot [2i\sin(\pi\nu\alpha)]^\tau \\
&= \left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^\tau e^{-\pi i\nu\alpha\varphi\cdot\tau}
\end{aligned} \tag{A.26}$$

Multiplikation mit $e^{2\pi i\nu(y-x)}$ ergibt hieraus für den Realteil

$$Re[(\tilde{f}_{(L,R)}(2\pi i\nu))^{\tau} e^{2\pi i\nu(y-x)}] = \left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau} Re(e^{-\pi i\nu\alpha\varphi\cdot\tau+2\pi i\nu(y-x)}) \qquad (A.27)$$
$$= \left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau} \cos(\pi\nu\alpha\varphi\tau - 2\pi\nu(y-x))$$

Wir haben jetzt also den Realteil des zentralen Terms in der Gleichung (A.19) für die bedingte Dichte der Hintergrundsequenz $Y_\tau$ für den Spezialfall einer (L,R)-parametrisierten Innovationssequenz hergeleitet und können diesen Ausdruck daher in (A.19) einsetzen:

$$Re\ g_\tau(y|x) = 1 + 2\sum_{\nu=1}^{\infty}\left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau}\cos(\pi\nu\alpha\varphi\tau - 2\pi\nu(y-x)) \qquad (A.28)$$

Nachdem wir vorab diesen Ausdruck für $Re\ g_\tau(y|x)$ berechnet haben, betrachten wir nun eine unverzerrte Hintergrundsequenz $Y_\tau$, die sich von der entsprechenden $(\alpha, \varphi)$-parametrisierten Innovationssequenz ableitet. Die Autokorrelationsfunktion von $Y_\tau$ ist definiert als

$$R_Y(\tau) = \frac{E[(Y_\tau - \mu_Y)(Y_0 - \mu_Y)]}{\sigma_Y^2} = \frac{E[Y_\tau Y_0] - \mu_Y^2}{\sigma_Y^2} \qquad (A.29)$$

mit $\mu_Y$ als Erwartungswert und $\sigma_Y^2$ als Varianz der Zufallsvariablen $Y$. Jetzt können wir die Tatsache ausnützen, daß sich die Übergangsdichte $g_\tau(y|x)$ des ursprünglich betrachteten Markovprozesses auch im Sinne einer gemeinsamen Verteilung von $Y_\tau$ und $Y_0$ interpretieren läßt (d.h. als die Wahrscheinlichkeit dafür, daß $Y_\tau = y$ und gleichzeitig $Y_0 = x$ gilt, was offensichtlich äquivalent zur Wahrscheinlichkeit für $Y_\tau = y$ unter der Bedingung $Y_0 = x$ ist). Hieraus ergibt sich

$$R_Y(\tau) = \frac{E[Y_\tau Y_0] - \mu_Y^2}{\sigma_Y^2} = \frac{\int_0^1\int_0^1 xy\cdot g_\tau(y|x)dxdy - \mu_Y^2}{\sigma_Y^2} = \qquad (A.30)$$

$$= \frac{\int_0^1\int_0^1 xy\cdot\left(1 + 2\sum_{\nu=1}^{\infty}\left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau}\cos(\pi\nu\alpha\varphi\tau - 2\pi\nu(y-x))\right)dxdy - \mu_Y^2}{\sigma_Y^2} =$$

$$= \frac{\int_0^1\int_0^1 xydxdy + 2\sum_{\nu=1}^{\infty}\left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau}\int_0^1\int_0^1 xy\cos(\pi\nu\alpha\varphi\tau - 2\pi\nu(y-x))dxdy - \mu_Y^2}{\sigma_Y^2}$$

Mit

$$\int_0^1\int_0^1 xy\,dx\,dy \;=\; \left(\int_0^1 x\,dx\right)^2 \;=\; \mu_Y^2 \;=\; \frac{1}{4} \tag{A.31}$$

erhält man daraus schließlich

$$R_Y(\tau)= \frac{2\sum_{\nu=1}^{\infty}\left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau}\int_0^1\int_0^1 xy\cos(\pi\nu\alpha\varphi\tau-2\pi\nu(y-x))\,dx\,dy}{\sigma_Y^2}\;. \tag{A.32}$$

Letzteres Integral läßt sich z.B. mit Mathematica lösen und ergibt

$$\int_0^1\int_0^1 xy\cos(\pi\nu\alpha\varphi\tau-2\pi\nu(y-x)) \;= \tag{A.33}$$

$$= -\frac{-\cos\pi\nu\alpha\varphi\tau+\cos\pi\nu(\alpha\varphi\tau-2)+2\pi\nu\sin\pi\nu(\alpha\varphi\tau-2)}{16\pi^4\nu^4}+$$

$$+\frac{\cos\pi\nu\alpha\varphi\tau+4\pi^2\nu^2\cos\pi\nu\alpha\varphi\tau-\cos\pi\nu(\alpha\varphi\tau+2)-2\pi\nu\sin\pi\nu(\alpha\varphi\tau+2)}{16\pi^4\nu^4}$$

Unter Verwendung der Identitäten $\cos\pi\nu(\alpha\varphi\tau\pm 2) = \cos(\pi\nu\alpha\varphi\tau\pm 2\pi\nu) = \cos\pi\nu\alpha\varphi\tau$ und $\sin\pi\nu(\alpha\varphi\tau\pm 2) = \sin\pi\nu\alpha\varphi\tau$ ([BS87] S. 181) läßt sich dies weiter reduzieren auf

$$\int_0^1\int_0^1 xy\cos(\pi\nu\alpha\varphi\tau-2\pi\nu(y-x)) \;=$$

$$= -\frac{-\cos\pi\nu\alpha\varphi\tau+\cos\pi\nu\alpha\varphi\tau+2\pi\nu\sin\pi\nu\alpha\varphi\tau}{16\pi^4\nu^4}+$$

$$+\frac{\cos\pi\nu\alpha\varphi\tau+4\pi^2\nu^2\cos\pi\nu\alpha\varphi\tau-\cos\pi\nu\alpha\varphi\tau-2\pi\nu\sin\pi\nu\alpha\varphi\tau}{16\pi^4\nu^4}$$

$$= \frac{4\pi^2\nu^2\cos\pi\nu\alpha\varphi\tau}{16\pi^4\nu^4} \;=\; \frac{\cos\pi\nu\alpha\varphi\tau}{4\pi^2\nu^2} \tag{A.34}$$

Einsetzen dieses Ergebnisses in die ursprüngliche Gleichung (A.32) liefert schließlich

$$R_Y(\tau) = \frac{2\sum_{\nu=1}^{\infty}\left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau}\cdot\frac{\cos\pi\nu\alpha\varphi\tau}{4\pi^2\nu^2}}{\sigma_Y^2} = \frac{1}{\sigma_Y^2}\sum_{\nu=1}^{\infty}\left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau}\cdot\frac{\cos\pi\nu\alpha\varphi\tau}{2\pi^2\nu^2} \qquad \text{(A.35)}$$

Fassen wir an dieser Stelle zusammen: Die Interpretation der Hintergrundsequenz $Y_\tau$ (A.12) eines unverzerrten TES-Modelles als stationärer Markovprozeß erlaubt die Angabe der entsprechenden Übergangsdichte $g_\tau(y|x)$ (A.19). Zentraler Term hierbei ist die Laplacetransformierte der $\tau$-fachen Faltung der Innovationsdichte $f_V$ mit sich selbst, die im folgenden für den einfachen Fall einer (L,R)-parametrisierten Innovationssequenz berechnet wurde und in Gleichung (A.28) resultierte. Andererseits läßt sich $g_\tau(y|x)$ auch als Dichte der gemeinsamen Verteilung von $Y_\tau$ und $Y_0$ interpretieren und als solche direkt in die Definition der Autokorrelationsfunktion (A.29) einsetzen, was (A.32) und nach weiterer Vereinfachung schließlich den Ausdruck (A.34) ergibt.

Damit haben wir in einem sehr einfachen Fall die analytische Form der Autokorrelationsfunktion hergeleitet und wollen sie abschließend noch mit der in Lemma A.8 angegebenen Formel vergleichen. Da wir ein unverzerrtes Modell betrachteten, ist die entsprechende Distortion-Funktion $D$ die Identität $D(x) = x$, die in [JM92b] als Spezialfall untersucht wird, für dessen Laplacetransformierte gilt:

$$\left|\tilde{D}(2\pi i\nu)\right|^2 = \frac{1}{(2\pi\nu)^2}. \qquad \text{(A.36)}$$

Setzt man dies in (A.11) ein, so ergibt sich zusammen mit (A.26)

$$\rho_Y(\tau) = \frac{2}{\sigma_Y^2}\sum_{\nu=1}^{\infty} Re[\tilde{f}_V^{\tau}(\nu)]\left|\tilde{D}(\nu)\right|^2 = \frac{2}{\sigma_Y^2}\sum_{\nu=1}^{\infty}\left[\frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha}\right]^{\tau} e^{-\pi i\nu\alpha\varphi\cdot\tau}\frac{1}{(2\pi\nu)^2} \qquad \text{(A.37)}$$

was konsistent zum direkten Resultat von (A.35) ist.

Für weitere Details und Rechenbeispiele für die analytische Darstellung der Autokorrelationsfunktion in der allgemeinen Form von Lemma A.8 sei nochmals auf [JM92a] und [JM92b] sowie die in [Moh99] davon gegebene Zusammenfassung verwiesen.

## A.3 Beispiel für den Automatisierungsalgorithmus

Zur Illustration des in Abschnitt 4.4.2 angegebenen Automatisierungsalgorithmus' wird im folgenden in Anlehnung an [Moh99] ein einfaches Beispiel dazu explizit durchgerechnet. Sei hierzu eine Meßreihe mit der Periodenlänge 16 gegeben, die in zwei Klassen mit der Klassenbreite 30 aufgeteilt wird. Aus Gründen der Übersichtlichkeit geben wir nur die folgende "typische" Periode an:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 5 | 24 | 35 | 47 | 58 | 55 | 51 | 49 | 49 | 52 | 53 | 48 | 39 | 31 | 12 | 7 |

*1. Schritt:*
Die empirische Verteilungsfunktion hat folgende Form:

$$F(x) = \begin{cases} 0, \text{ falls } x < 0 \\ \frac{1}{120}x, \text{ falls } x \in [0, 30) \\ \frac{1}{40}x - \frac{1}{2}, \text{ falls } x \in [30, 60) \\ 1, \text{ falls } x \geq 60 \end{cases}. \tag{A.38}$$

*2. Schritt:*
Zur Schätzung der Innovationsverteilung wird die angegebene typische Periode verwendet.

*3. und 4. Schritt:*
Die Aufteilung in Monotonieintervalle ergibt die Werte $k = 3$, $n_1 = 5$, $n_2 = 8$ und $n_3 = 11$. Damit muß für die stückweise lineare Stitching-Funktion gelten:

$$m_1 \in \left(\frac{5}{17}, \frac{6}{17}\right), m_2 \in \left(\frac{8}{17}, \frac{9}{17}\right), m_3 \in \left(\frac{11}{17}, \frac{12}{17}\right),$$
$$S(m_1) \geq \frac{58}{60}, S(m_2) \leq \frac{49}{60} \text{ und } S(m_3) \geq \frac{53}{60}$$

Die Wahl der Punkte $(0.3, 1)$, $(0.5, 0.8)$ und $(0.7, 0.9)$ genügt diesen Bedingungen. Damit kann die Verallgemeinerte Stitching-Funktion (A.5) und deren Verteilungsfunktion aus Beispiel A.5 verwendet werden.

*5. Schritt:*

$$F_S^{-1}(x) = \begin{cases} \frac{30}{19}x, \text{ falls } x \in \left[0, \frac{38}{75}\right) \\ \frac{30}{109}x + \frac{72}{109}, \text{ falls } x \in \left[\frac{38}{75}, \frac{87}{100}\right) \\ \frac{10}{13}x + \frac{3}{13}, \text{ falls } x \in \left[\frac{87}{100}, 1\right] \end{cases}$$

| $x_i$ | 5 | 24 | 35 | 47 | 58 | 55 | 51 | 49 |
|---|---|---|---|---|---|---|---|---|
| $F(x_i)$ | $\frac{1}{24}$ | $\frac{1}{5}$ | $\frac{3}{8}$ | $\frac{27}{40}$ | $\frac{19}{20}$ | $\frac{7}{8}$ | $\frac{31}{40}$ | $\frac{29}{40}$ |
| $F_S^{-1}(F(x_i))$ | $\frac{5}{76}$ | $\frac{6}{19}$ | $\frac{45}{76}$ | $\frac{369}{436}$ | $\frac{25}{26}$ | $\frac{47}{52}$ | $\frac{381}{436}$ | $\frac{375}{436}$ |

| $x_i$ | 49 | 52 | 53 | 48 | 39 | 31 | 12 | 7 |
|---|---|---|---|---|---|---|---|---|
| $F(x_i)$ | $\frac{29}{40}$ | $\frac{4}{5}$ | $\frac{33}{40}$ | $\frac{7}{10}$ | $\frac{19}{40}$ | $\frac{11}{40}$ | $\frac{1}{10}$ | $\frac{7}{120}$ |
| $F_S^{-1}(F(x_i))$ | $\frac{375}{436}$ | $\frac{96}{109}$ | $\frac{387}{436}$ | $\frac{93}{109}$ | $\frac{3}{4}$ | $\frac{33}{76}$ | $\frac{3}{19}$ | $\frac{7}{76}$ |

*6. Schritt:*

$$S_1 = S|_{[0,\,0.3]},\ S_2 = S|_{[0.3,\,0.5]},\ S_3 = S|_{[0.5,\,0.7]} \text{ und } S_4 = S|_{[0.7,\,1]}$$

$$S_1^{-1}(x) = \frac{3}{10}x, \text{ für } x \in [0, 1],\ S_2^{-1}(x) = -x + \frac{13}{10}, \text{ für } x \in [0.8, 1],$$

$$S_3^{-1}(x) = 2x + \frac{11}{10}, \text{ für } x \in [0.8, 0.9] \text{ und } S_4^{-1}(x) = -\frac{1}{3}x + 1, \text{ für } x \in [0, 0.9]$$

*7. Schritt*:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $u_i$ | 0.0197 | 0.0947 | 0.1776 | 0.2539 | 0.2885 | 0.3962 | 0.4261 | 0.4399 |

| $i$ | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|
| $u_i$ | 0.6202 | 0.6614 | 0.6752 | 0.7156 | 0.75 | 0.8553 | 0.9474 | 0.9693 |

*8. Schritt*:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $v_i$ | 0.075 | 0.0829 | 0.0763 | 0.0346 | 0.1077 | 0.0299 | 0.0138 | 0.1803 |

| $i$ | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|
| $v_i$ | 0.0412 | 0.0138 | 0.0404 | 0.0344 | 0.1053 | 0.0921 | 0.0219 |

Also:

$$\hat{\mu}_V = \frac{0.9496}{15} = 0.0633 \text{ und } \hat{\sigma}_V^2 = \frac{0.0293}{14} = 0.002096$$

*9. Schritt:*

$$\alpha = \frac{0.0633^2}{0.002096} = 1.911 \text{ und } \lambda = \frac{0.0633}{0.002096} = 30.2$$

Damit sind aus der gegebenen Meßreihe Innovationsvariablen ermittelt worden, die $\Gamma(1.991;30.2)$-verteilt sind. Eine Rundung des Parameters $\alpha$ auf einen ganzzahligen Wert vereinfacht die Implementierung des Modellierungsverfahrens, da die Innovationsverteilung dann einer Erlang-$\alpha$-Verteilung mit Parameter $\lambda = 30.2$ entspricht.

# A.4 Erweiterung der Refined Uniform Asymptotic Approximation

In Abschnitt 7.2.4 wurde erwähnt, daß nach [MRM98] noch eine weitergehende Approximation von $K$ gemäß (7.32) bzw. (7.33) existiert, die numerisch korrekte Ergebnisse unabhängig davon liefert, ob $z^*$ nahe bei 1 liegt oder nicht. Hierzu sei

$$\Gamma^{(1)} = \begin{cases} \dfrac{z^* \log z^* + 1 - z^*}{(1-z^*)^2}, \forall(|1-z^*| > \varepsilon_1) \\ \displaystyle\sum_{i=0}^{\infty} \frac{(1-z^*)^i}{(i+1)(i+2)}, \forall(|1-z^*| \le \varepsilon_1) \end{cases} \tag{A.39}$$

$$\Gamma^{(2)} = \begin{cases} \dfrac{2{z^*}^2 \log z^* + 2z^*(1-z^*) - (1-z^*)^2}{(1-z^*)^3}, \forall(|1-z^*| > \varepsilon_2) \\ -4\displaystyle\sum_{i=0}^{\infty} \frac{(1-z^*)^i}{(i+1)(i+2)(i+3)}, \forall(|1-z^*| \le \varepsilon_2) \end{cases} \tag{A.40}$$

Hierbei garantiert die Wahl von $\varepsilon_1 = \varepsilon_2 = 0.1$ und die Approximation der unendlichen Summe durch ihre ersten 12 Summanden numerisch brauchbare Resultate

Sei weiterhin

$$\phi = \frac{\displaystyle\sum_{s=1}^{S} \frac{\nu_s}{C} \sum_{n=0}^{ds-1} n(z^*)^n + \Gamma^{(1)}}{z^*} \tag{A.41}$$

und

$$\Psi = \sum_{s=1}^{S} \frac{\nu_s}{C} \sum_{n=0}^{d_s - 1} n(n-1)(z^*)^n + \Gamma^{(2)} \tag{A.42}$$

Dann lassen sich die Gleichungen (7.32) und (7.33) ersetzen durch

$$K = 1 + \frac{\Psi}{\sqrt{2\phi}(z^*\sqrt{2\phi} + \sqrt{V})} \tag{A.43}$$

Auf ähnliche Weise kann man auch einen Ausdruck für $E$ herleiten, der (7.41) und (7.42) ersetzt und unabhängig davon ist, ob $z^*$ nahe bei 1 liegt oder nicht. Sei

$$\Gamma^{(3)} = \begin{cases} \dfrac{6z^{*3}\log z^* + 6z^{*2}(1-z^*) - 3z^*(1-z^*)^2 + 2(1-z^*)^3}{(1-z^*)^4}, & \forall(|1-z^*| > \varepsilon_3) \\ 36\displaystyle\sum_{i=0}^{\infty} \frac{(1-z^*)^i}{(i+1)(i+2)(i+3)(i+4))}, & \forall(|1-z^*| \le \varepsilon_3) \end{cases} \tag{A.44}$$

und

$$\Gamma^{(4)} = \begin{cases} \dfrac{24z^{*4}\log z^* + 24z^{*3}(1-z^*) - 12z^{*2}(1-z^*)^2 + 8z^*(1-z^*)^3 - 6(1-z^*)^4}{(1-z^*)^5} & \\ \qquad\qquad \forall(|1-z^*| > \varepsilon_4) & \\ -576\displaystyle\sum_{i=0}^{\infty} \frac{(1-z^*)^i}{(i+1)(i+2)(i+3)(i+4)(i+5)}, & \forall(|1-z^*| \le \varepsilon_4) \end{cases} \tag{A.45}$$

Numerisch brauchbare Ergebnisse erhält man hier für $\varepsilon_3 = 0.2$ und $\varepsilon_4 = 0.275$, die unendlichen Summen sind für $\Gamma^{(3)}$ bis zum 15. Glied und für $\Gamma^{(4)}$ bis zum 18. Glied zu berechnen.

Sei weiterhin

$$\chi = \sum_{s=1}^{S} \frac{\nu_s}{C} \sum_{n=0}^{d_s-1} n(n-1)(n-2)(z^*)^n + \Gamma^{(3)} \tag{A.46}$$

und

$$\theta = \sum_{s=1}^{S} \frac{\eta_s}{C} \sum_{n=0}^{d_s-1} n(n-1)(n-2)(n-3)(z^*)^n + \Gamma^{(4)} \tag{A.47}$$

Dann ergibt sich die verbesserte uniforme Approximation von $E$ zu

$$\begin{aligned} E = {} & \frac{\theta + Y}{8z^*V} + K^3 - \frac{9z^*\Psi}{8V^2} - \frac{\chi}{V}\left(1 + \frac{5(T + 3z^*\Psi)}{24V}\right) \\ & + \frac{(3)\Psi^2(K-1)[3(z^*\sqrt{2\phi})^2 + 9\sqrt{V}(z^*\sqrt{2\phi})^2 + 8V]}{8V^2(z^*\sqrt{2\phi} + \sqrt{V})^2} \\ & + \frac{3\psi(z^*\sqrt{2\phi} + 2\sqrt{V})}{2V(z^*\sqrt{2\phi} + \sqrt{V})}(K-1)\left(K + 1 + \frac{z^*\psi}{2V}\right) \end{aligned} \tag{A.48}$$

ANHANG B

# *Zur Berücksichtigung von Nutzerpräferenzen bei der Dienstvermittlung in einem Trader*

Utility-Funktionen, wie wir sie in Abschnitt 8.2 im Zusammenhang mit Auktionen kennengelernt haben, spielen auch in anderen anwendungsbezogenen Kontexten eine wichtige Rolle. Im folgenden wird exemplarisch dafür untersucht, wie sich durch Einführung einer derartigen Funktion in mehreren Dimensionen (unter dem Namen *QoS-Funktion*) der klassische Trading-Service unter CORBA dynamischer und flexibler gestalten läßt. Für detailliertere Untersuchungen hierzu sei auf [RTL96] und [RLT97] sowie auch [LRT97] verwiesen.

## B.1 Einführung

Dem zunehmenden Aufwand für die Entwicklung immer komplexer werdender verteilter Informationssysteme läßt sich am ehesten durch die Wiederverwendung bereits existierender Komponenten begegnen, wie dies objektorientierte Technologien zum Ziel haben. Die Verwendung von Middleware erlaubt es weiterhin, die Heterogenität von bestehenden Rechnern unterschiedlicher Hersteller zu überbrücken und sie in ein gemeinsames Rechnernetz zu integrieren. Ein bekanntes Beispiel hierfür bildet die von der Object Management Group (OMG) standardisierte Common Object Request Broker Architecture (CORBA) (vgl. [corba], [PSW96], [Lin98]).

Kernstück von CORBA ist der Object Request Broker (ORB), der mit drei Klassen von Diensten in Verbindung steht: den Basisdiensten (CORBAservices), welche unabhängig von bestimmten Anwendungen im CORBA-Standard spezifiziert sind, den eher anwendungsorientierten Common Facilities und den Applikationsdiensten, die als Clients und Server vom Entwickler programmiert und vom Anwender direkt angesprochen werden.

Im folgenden betrachten wir einen der Basisdienste näher, und zwar den CORBA Trading Service ([Po95], [RTL96]) und gehen von einem unter dem CORBA-Produkt Orbix implementierten Trading Service aus [orbix96], wie er an der RWTH Aachen realisiert wurde [MZP96]. Bei

der Dienstvermittlung dieses Traders kann ein Dienst nur dann angeboten werden, wenn er *genau* mit den vom Kunden angeforderten Charakteristiken übereinstimmt.

Diese Einschränkung wird durch die im folgenden vorgestellte Erweiterung aufgehoben, um damit die Berücksichtigung von Nutzerinteressen zu ermöglichen. Die Grundidee ist einfach: Der gewünschte Dienst wird neben dem Diensttyp noch über sogenannte Quality of Service (QoS)-Attribute charakterisiert, welche Leistungsmerkmale angebotener Dienste beschreiben [qos95]. Der Kunde des Systems erhält dann die Möglichkeit, quantitativ anzugeben, welche QoS-Attribute er wie präferieren möchte. Daraufhin wird ihm der *am besten geeignete* Dienst empfohlen, bevor er selbst entscheiden kann, ob er diesen in Anspruch nehmen will oder nicht.

Im nächsten Abschnitt wird auf die grundlegenden Konzepte des unter Orbix implementierten Traders eingegangen, während Abschnitt B.3 ein Konzept zur Berücksichtigung von Nutzerinteressen für die Auswahlentscheidung des Traders skizziert. Abschnitt B.4 gibt dann einen Überblick über die Modellierung der Entscheidungskomponente. Schließlich wird in Abschnitt B.5 kurz auf die Implementierung dieser Ansätze eingegangen, bevor einige Meßergebnisse vorgestellt werden.

## B.2 Trading im Kontext von CORBA

Abbildung B-1 stellt die CORBA zugrundeliegende Architektur dar. Für die Dienstvermittlung exportieren die Serverobjekte ihre Objektreferenz und geben zur Beschreibung der Semantik ihres Dienstes den Diensttyp und zugehörige Dienstattribute an, welche die Qualität des Dienstes im Sinne eines QoS-Merkmals beschreiben. Sucht ein Kunde nach einem Dienst bzw. benötigt ein Objekt eine entsprechende Referenz zur Laufzeit des Systems, so vermittelt der Trading Service die Objektreferenz, und der als geeignet ausgewählte Dienst wird dynamisch gebunden. Liegen Client- und Serverobjekte auf heterogenen verteilten Rechnern vor, so wird die Funktionalität des ORBs ausgenutzt, um Interoperabilität zwischen den verschiedenen Objekten zu ermöglichen.

Dieses Prinzip des Tradings ist bereits seit einigen Jahren bekannt [SPM94]. Es beruht darauf, daß der dienstsuchende Kunde die gewünschten Eigenschaften des benötigten Dienstes exakt spezifiziert. Die Struktur des Dienstes läßt sich dadurch charakterisieren, daß zu einem Diensttyp verschiedene Dienstangebote existieren können. Diese Dienstangebote sind die konkrete Beschreibung der innerhalb des CORBA-Systems von den Objekten angebotenen Funktionalitäten. Sie werden durch Eigenschaften oder Attribute in Form von (Name, Typ, Wert)-Tripeln charakterisiert. Wichtig ist dabei lediglich, daß im Service Directory, auf das die Trading-Funktion zugreift, Dienste vorhanden sind, welche die Anforderungen des Kunden erfüllen. Dazu muß der geforderte mit einem vorhandenen Diensttyp übereinstimmen und die Menge der spezifizierten QoS-Eigenschaften die Restriktionen der Anfrage erfüllen, wie Abbildung B-2 verdeutlicht.

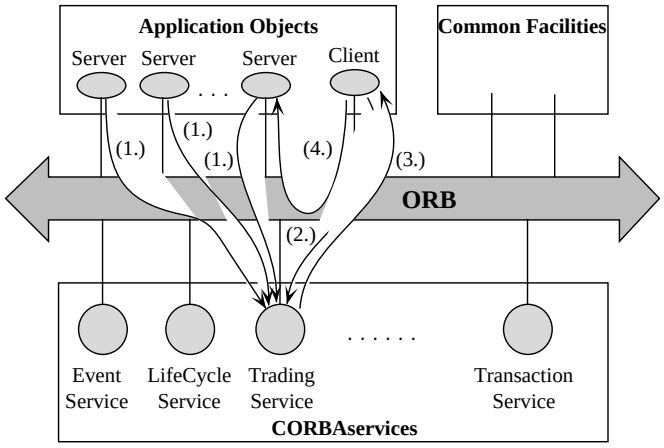


**Abbildung B-1:** Das Prinzip des Tradings in einer CORBA-Umgebung

Diese standardisierte Trading-Funktion wird nun erweitert, um dem Nutzer mehr Möglichkeiten für die Dienstsuche oder Dienstauswahl zur Verfügung zu stellen. Übernommen wird diese Funktionalität durch die im folgenden definierte *Quality-of-Service-Funktion*, die die numerische Bewertung eines Dienstangebots in Abhängigkeit von den Interessen des Nutzers erlaubt. Diese Funktion läßt sich zusammensetzen aus sogenannten *Quality-of-Service-Property-Funktionen*, deren jede die Charakterisierung einer einzelnen Dienst*eigenschaft* des entsprechenden Dienstes hinsichtlich der Nutzerpräferenzen ermöglicht. Über das in [Thi96] und [TP96] entwickelte Konzept eines Abstands zwischen Dienstanfrage und -angeboten hinausgehend werden dabei die Präferenzen des Nutzers unter Verwendung von Konzepten der präskriptiven Entscheidungstheorie (s. [KR76], [EW93]) miteinbezogen. Dabei wird vorausgesetzt, daß die in einem Dienstangebot spezifizierten Attribute im Falle einer Annahme auch tatsächlich eingehalten werden, während in [MLB93] mit der sogenannten Agency Theory eine Idee skizziert wird, wie man diesbezügliche "Unsicherheiten" modellieren könnte.

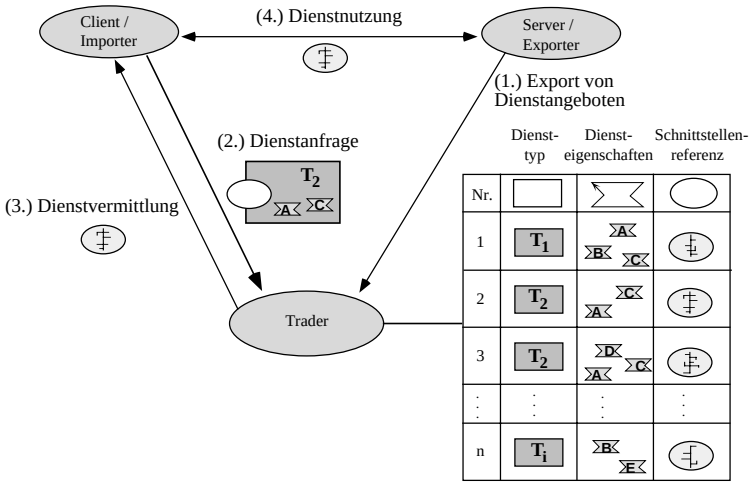


**Abbildung B-2:** Schnittstellenreferenzen angebotener und nachgefragter Dienste

## B.3 QoS- und QoSP-Funktionen

Nutzerpräferenzen hinsichtlich der Eigenschaften eines Dienstes lassen sich grundsätzlich in zwei Richtungen unterteilen. Zum einen kann der Nutzer festlegen, wie wichtig ihm eine Diensteigenschaft als ganze verglichen mit anderen Diensteigenschaften ist (Gewichtung der Diensteigenschaften). Zum anderen kann der Nutzer aber auch Präferenzen in Bezug auf jede einzelne Diensteigenschaft haben, indem bestimmte Ausprägungen des jeweiligen Attributes für ihn einen bestimmten Wert besitzen. Dies läßt sich durch eine für die betrachtete Diensteigenschaft spezifische Funktion charakterisieren, die jeder Attributausprägung eine Zahl zwischen 0 und 1 zuordnet, um den nutzerspezifischen Wert auszudrücken, den diese Ausprägung aufweist.

Abstrakter ausgedrückt wird also zur Berücksichtigung der Nutzerpräferenzen für jeden Dienst eine Quality-of-Service-Funktion (*QoS-Funktion*) eingeführt, anhand derer jedes Dienstangebot vom Trader bewertet wird. Die QoS-Funktion eines Dienstes setzt sich zusammen aus Bewertungsfunktionen für jede einzelne Diensteigenschaft, den Quality-of-Service-Property-Funktionen (*QoSP-Funktionen*).

Die Bestimmung dieser Funktionen erfolgt durch den importierenden Nutzer im Dialog mit einer *Assessment-Komponente*, und zwar bevor eine Dienstanfrage an den Trader gestellt werden kann. Die Anfrage beinhaltet dann neben den üblichen Einträgen zu jeder angefragten Diensteigenschaft $d$ die entsprechende QoSP-Funktion $P_d$ sowie ein Gewicht $w_d$. Aus diesen Angaben kann der Trader dann in einer neuen Evaluator-Komponente die verschiedenen ihm vorliegenden Dienstangebote nutzerspezifisch bewerten. Abbildung B-3 verdeutlicht (im Schritt 2a) das Vorschalten der Assessment-Komponente vor die eigentliche Anfrage und die hieraus resultierende Bewertung der einzelnen Dienstangebote durch die Evaluator-Komponente des erweiterten Traders.

Dabei ist zwischen zwei verschiedenen Typen von Diensteigenschaften zu unterscheiden: Zum einen gibt es Attribute, für die der Nutzer genau einen Wert bzw. obere oder untere Schranken angeben kann, die dieses Attribut in jedem Fall einhalten muß. So sind etwa für den Ausdruck eines DIN-A3-Blatts alle Drucker, die lediglich DIN-A4 anbieten können, schlichtweg nutzlos. Solche "K.o.-Kriterien" lassen sich mit QoSP-Funktionen behandeln, die außerhalb des Zielwertes bzw. der Schranken identisch 0 sind. Man kann sie in der Anfrage aber auch wie bisher z.B. über die "conditional-selection-criteria" der Anfragesprache SRDL (Service Request Description Language, vgl. [PM95]) formulieren. Der zweite Typ von Diensteigenschaften besteht aus den "evaluation-criteria", zwischen denen je nach Interessenlage des Nutzers ein Kompromiß gefunden werden muß. Damit kann man eine Anfrage mit Hilfe einer erweiterten Version von SRDL folgendermaßen formulieren:

**"SELECT"** <service-type-identifier> **"WITH"**
"IF SUCCESS"

<conditional-selection-criteria>
**"THEN"**
<evaluation-criteria>
**"ELSE END"**

evaluation-criteria ::=<service-property-identifier> **"WITH VALUE FUNCTION"** <value-vector> **"AND WEIGHT"** <weight>
|<evaluation-criteria> **"AND"** <evaluation-criteria>

value-vector ::=<value-point> | <value-vector> **"AND"** <value-vector>
value-point ::="**(**" <real-number> "**,**" <real-number> "**)**"
weight ::=<real-number>

Im folgenden Abschnitt wird nun genauer darauf eingegangen, woher man die zu übergebenden Anfragevariablen value-vector und weight erhält.

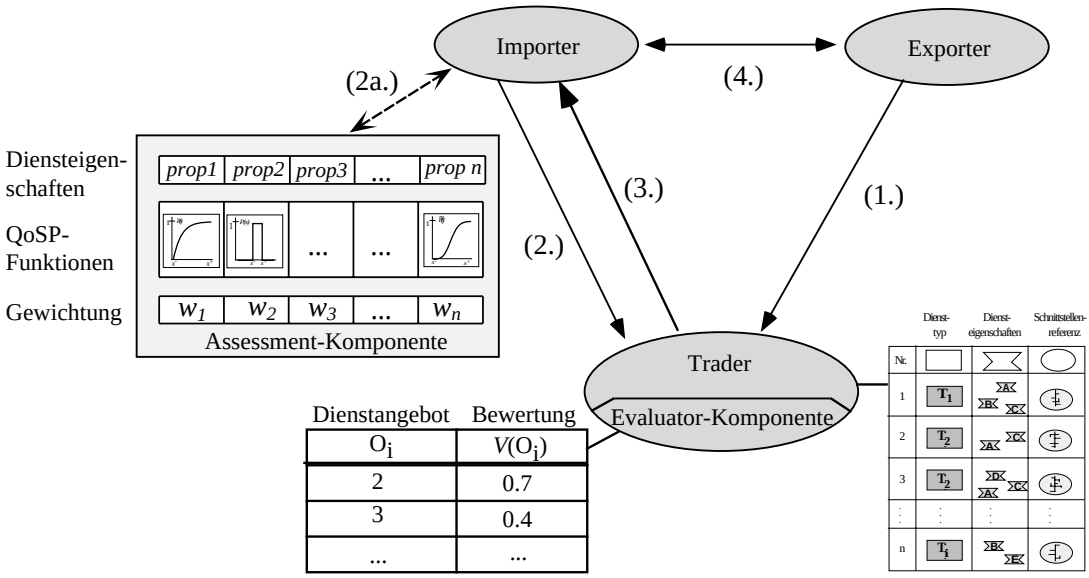


**Abbildung B-3:** Assessment-Komponente und Evaluator-Komponente

# B.4 Modellierung von Nutzerinteressen

Um eine Entscheidung zwischen verschiedenen Dienstangeboten treffen zu können, werden diese mit Hilfe der QoS-Funktion bewertet; die Entscheidung fällt dann auf das Angebot mit der höchsten Bewertung. Die QoS-Funktion hängt dabei von den einzelnen Diensteigenschaften und deren von Nutzer zu Nutzer unterschiedlichen Gewichtung ab. Außerdem ist zu berücksichtigen, daß auch die möglichen Ausprägungen der einzelnen Diensteigenschaften eine nutzerspezifische Bewertung besitzen.

Betrachten wir zur Veranschaulichung einen Dienst vom Typ *Printer*, der z.B. Druckgeschwindigkeit, Auflösung und räumliche Entfernung des Geräts als Diensteigenschaften aufweist. Dann sind Nutzer denkbar, die viel Zeit haben und denen es daher gleichgültig ist, wie schnell ein Druckvorgang erfolgt. Ein solcher Nutzer wird die verschiedenen möglichen Ausprägungen des Attributs Druckgeschwindigkeit gleich hoch bewerten, während ein eiliger Zeitgenosse eine hohe Druckgeschwindigkeit einer niedrigen vorziehen wird. Darüber hinaus wird jeder Nutzer individuelle Vorstellungen haben, wie wichtig ihm die Druckgeschwindigkeit als solche im Vergleich etwa zur räumlichen Entfernung des Gerätes sein wird.

Hieraus ergibt sich ein Vorgehen in zwei Schritten: Zunächst muß der Nutzer für jede relevante Diensteigenschaft $d$ eine QoSP-Funktion $P_d$ angeben. Im zweiten Schritt ist dann noch die Gewichtung der einzelnen QoSP-Funktionen zu ermitteln, mit der sie in die QoS-Funktion $V$ eingehen sollen. Für beide Schritte werden im folgenden einige Verfahren beschrieben.

### B.4.1 Ermittlung einer QoSP-Funktion

Ausprägungen einer Diensteigenschaft lassen sich mit Hilfe geeigneter *Attribute* darstellen. Typische Beispiele hierfür sind etwa die Geschwindigkeit eines Druckers mit dem Attribut "Seiten pro Minute" oder die Kapazität einer Netzverbindung in Mbps. Solche Attribute können grundsätzlich *quantitativ,* d.h. in Form einer ganzen oder reellen Zahl, oder *qualitativ,* d.h. in Form von Noten wie "sehr gut", "ausreichend", "ungenügend" etc., vorliegen.

Betrachten wir zunächst ein quantitatives Attribut $x$. Dann läßt sich i.d.R. eine untere Grenze $x^-$ und eine obere Grenze $x^+$ angeben, zwischen denen sämtliche möglichen Attributwerte liegen. Der Nutzer wird nun verschiedenen Attributwerten $x_i$ unterschiedliche Bewertungen $P(x_i)$ zukommen lassen. Normiert man den Wertebereich der Funktion $P$ auf das Intervall [0;1], und nehmen wir $P(x^-) = 0$ und $P(x^+) = 1$ an, d.h. $x^-$ wird am niedrigsten und $x^+$ am höchsten bewertet, so ergibt sich typischerweise ein Kurvenverlauf wie in Abbildung B-4a. Gibt der Nutzer eine Schwelle vor, die eine Diensteigenschaft überschreiten muß, so entspricht das einer stufenförmigen Funktion, analog läßt sich die Vorgabe einer unteren und oberen Schranke durch eine Rechteckfunktion darstellen (Abbildung B-4 (b)). Der Fall eines vom Nutzer vorgegebenen Zielwertes schließlich läßt sich näherungsweise durch eine Zackenfunktion wie in Abbildung B-4 (c) darstellen.

Für die Ermittlung von Funktionen zur Bewertung einzelner Diensteigenschaften durch den Nutzer greifen wir auf zwei weitverbreitete Verfahren zurück (vgl. [EW93], [KR76]): das *Direct Rating* und die *Midvalue Splitting Technique*. Die Verwendung des Direct Rating entspricht der "Benotung" einiger ausgesuchter Ausprägungen des Attributs, während bei der Midvalue Splitting Technique die zu einzelnen "Noten" (also den jeweiligen Werten von $P$) gehörigen Attributausprägungen bestimmt werden. Damit lassen sich QoSP-Funktionen mit quantitativen wie auch qualitativen Attributen angeben. Eine ausführlichere Gegenüberstellung der beiden Verfahren findet sich in [RLT97].

.

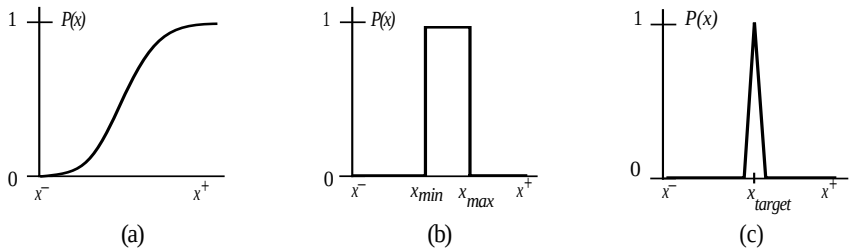


**Abbildung B-4:** Mögliche Kurvenverläufe einer QoSP-Funktion

### Direct Rating

Beim Direct Rating (Abbildung B-5 (a)) wählt man einige typische Ausprägungen $x_i$, die das entsprechende Attribut annehmen kann, und bewertet sie direkt mit einer Zahl $y_i$ zwischen 0 und 1. Für diese Benotung kann man z.B. das aus der Schule bekannte sechsstufige Notensystem zugrunde legen (mit den Noten "sehr gut" = 1, "gut" = 0.8, "befriedigend" = 0.6, "ausreichend" = 0.4, "mangelhaft" = 0.2 und "ungenügend" = 0). Durch lineare Interpolation der so gewonnenen Stützstellen erhält man eine hinreichend gute Approximation der QoSP-Funktion (Abbildung B-5 (b)).

### Midvalue Splitting Technique

Dieser Ansatz ist etwas umständlicher, kann aber dafür den Anspruch größerer "Objektivität" erheben. Hierbei werden zunächst $x^0 = x^-$ und $x^1 = x^+$ als die am schlechtesten bzw. am besten bewertete Attributausprägung ermittelt. In einem zweiten Schritt ist der "wertmäßige Mittelpunkt" $x^{0.5}$ des Intervalls $[x^0; x^1]$ anzugeben (s. Abbildung 5 (c)); hierunter versteht man diejenige Attributausprägung, für die der *Übergang* $x^0 \rightarrow x^{0.5}$ vom Nutzer genauso hoch bewertet wird wie der Übergang $x^0 \rightarrow x^{0.5}$. Setzt man also $P(x^0) = 0$ und $P(x^1) = 1$, so ergibt sich $P(x^{0.5}) = 0.5$. Analog dazu kann man noch $x^{0.25}$ und $x^{0.75}$ ermitteln und die gewonnenen fünf Stützstellen wiederum durch lineare Interpolation zu einer QoSP-Funktion $P$ fortsetzen.

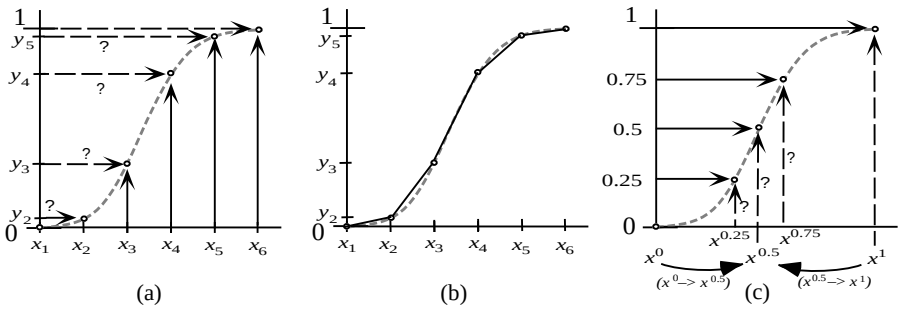


**Abbildung B-5:** Techniken zur Ermittlung der QoSP-Funktionen: (a) Direct Rating (b) Lineare Interpolation (c) Halbierungsmethode

## B.4.2 Ermittlung der QoS-Funktion

Im letzten Abschnitt haben wir Verfahren betrachtet, mit denen die Präferenzstruktur einer einzelnen Diensteigenschaft dargestellt werden kann. Wie bereits erwähnt, ist für die Auswahl eines Dienstes durch den Trader in der Regel aber mehr als eine Diensteigenschaft relevant. Daher müssen die alternativen Dienstangebote durch eine multiattributive Bewertungsfunktion eingeschätzt werden. Am einfachsten faßt man hierzu die verschiedenen QoSP-Funktionen mittels einer geeigneten Gewichtung zu einer linearen QoS-Funktion zusammen.

Betrachten wir also einen Dienst mit $n$ Diensteigenschaften. Wenn wir zu jeder davon die entsprechende QoSP-Funktion sowie das Gewicht $w_d$ kennen, das die jeweilige Eigenschaft im Vergleich zu den anderen Eigenschaften besitzt, dann ergibt sich die QoS-Funktion $V$ in Abhängigkeit vom $i$-ten Dienstangebot $O_i$ zu

$$V(O_i) = \sum_{d=1}^{n} w_d \cdot P_d(O_i) \tag{B.1}$$

### Ermittlung der Gewichte

Auch für die Ermittlung der Gewichte stehen mehrere etablierte Verfahren zur Verfügung, von denen hier kurz das Tradeoff- und das Swing-Verfahren vorgestellt werden sollen (vgl. [EW93], ein Vergleich beider Methoden findet sich in [RLT97]).

Das *Tradeoff*-Verfahren beruht auf der Ermittlung des Tradeoffs zwischen zwei Dienstattributen. Vereinfacht ausgedrückt wählt man dazu zwei Attribute aus, läßt eines davon eine schlechter bewertete Ausprägung annehmen und fragt danach, um wieviel das andere besser sein müßte, damit die Gesamtbewertung des Dienstes gleich bleibt, wobei alle übrigen Attribute im Vergleich zu vorher nicht verändert werden. Dies führt man für $n$-1 Alternativenpaare ($a$, $b$) durch, die sich jeweils nur in zwei Attributausprägungen unterscheiden und vom Nutzer gleich bewertet werden. Zusammen mit der Normierungsbedingung für die Gewichte ergeben diese Indifferenzaussagen ein $n$-dimensionales lineares Gleichungssystem, durch dessen eindeutige Lösung wir die gesuchten Gewichte erhalten.

Das *Swing*-Verfahren geht üblicherweise von einer "Alternative"

$$a^- = (a_1^-, a_2^-, \ldots, a_n^-) \tag{B.2}$$

aus, bei der alle Attribute jeweils den schlechtestmöglichen Wert annehmen. Relativ dazu werden sämtliche Alternativen der Form

$$b^r = (a_1^-, a_2^-, \ldots, a_{r-1}^-, a_r^+, a_{r-1}^-, \ldots, a_n^-) \tag{B.3}$$

bewertet, die mit Ausnahme des *r*-ten Attributes identisch zu $a^-$ sind, wobei das *r*-te Attribut nun aber seine am besten bewertete Ausprägung annimmt. Dies entspricht also der Bestimmung der Kosten, die der Nutzer zu investieren bereit ist, um Alternative $b^r$ anstatt Alternative $a^-$ zu erhalten. Aus den entstehenden *n* Gleichungen lassen sich unter Ausnutzung der Linearitätsannahme sofort die gesuchten Gewichte ableiten.

## B.5 Implementierung und Ergebnisse

In der vorgenommenen Implementierung der vorgestellten Ansätze erhält der Trader also eine Anfrage nach einem bestimmten Diensttyp, wobei er die "conditional-selection-criteria" so behandelt, wie er dies im herkömmlichen Fall gewohnt ist. Die danach noch verbleibenden Angebote müssen nun anhand der vom Nutzer angegebenen QoS- und QoSP-Funktionen evaluiert werden. Der Einfachheit halber gibt der Importer die QoSP-Funktionen für die einzelnen Dienstattribute anhand von Stützstellen an, zwischen denen dann die Evaluator-Komponente linear interpoliert. Mittels der ebenfalls vom Importer spezifizierten Gewichtung werden dann die Werte der QoS-Funktion aus (B.1) ermittelt, woraufhin der Trader dem Importer das am höchsten bewertete Angebot als optimal vorschlagen kann.

Die eigentliche Bewertung eines Dienstangebotes geht von einer sortierten Liste von Eigenschaften des Dienstangebotes aus. Diese kann auf zweierlei Weisen verwendet werden: entweder (Version A) werden Diensteigenschaften bzw. Dienstangebotseigenschaften mittels einer entsprechenden Funktion sortiert, oder aber (Version B) die nötige Sortierung wird schon beim Export eines Dienstes vorgenommen und dadurch bei größerem Speicheraufwand weniger zeitintensiv. Abbildung B-6 zeigt das Ergebnis eines Laufzeitvergleichs dieser Implementierung in den beiden beschriebenen Versionen A und B mit einem herkömmlichen Trader, wie er in [MZP96] vorgestellt wurde. Die Messung fand dabei auf einer SPARC 5 unter Verwendung von Orbix 2.0 bei geringer bis mittlerer Auslastung des Rechners statt. Als wesentliches Ergebnis kann man festhalten, daß sich der zeitliche Aufwand für die beschriebene Erweiterung der Funktionalität in durchaus vertretbaren Grenzen hält.

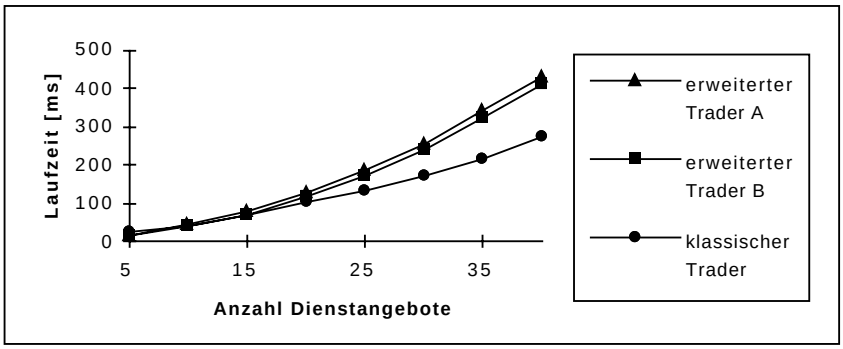


**Abbildung B-6:** Laufzeitvergleich der zwei Versionen des erweiterten Traders

Als zweiter Aspekt wurde noch untersucht, wie sich quantifizieren läßt, inwieweit der höhere Aufwand bei der Erstellung einer Anfrage dem Nutzer einen Vorteil bringt. Die Resultate hierzu sind in [RLT97] zusammengestellt.

## B.6 Fazit

Herkömmliche Trading Services unter CORBA bieten dem Nutzer nur dann einen Dienst an, wenn dieser genau die Diensteigenschaftsspezifikationen erfüllt, die der Nutzer in seiner Anfrage vorgibt. Das hier vorgestellte Konzept unterscheidet zwischen Anforderungen, die in jedem Fall zu erfüllen sind, und solchen, zwischen denen ein "Kompromiß" möglich ist. Hierzu ist es notwendig, daß der Nutzer in seiner Anfrage zusätzlich angibt, welches Gewicht er den einzelnen Diensteigenschaften beimißt und auf welche Weise er die einzelnen Ausprägungen jeder Diensteigenschaft präferiert. Es wurden ausführlich Methoden vorgestellt, um an diese Informationen zu gelangen. Sie ermöglichen es dem Trader, die einzelnen Dienstangebote nutzerspezifisch zu bewerten und dem Nutzer dasjenige Angebot vorzuschlagen, das individuell für ihn am besten ist. An der vorgestellten Implementierung läßt sich nachweisen, daß die hierfür benötigte zusätzliche Laufzeit nicht allzu groß ist. Andererseits zeigt sich, daß der durch die neue Funktionalität gewonnene Nutzen für den Anwender diesen zusätzlichen Aufwand durchaus rechtfertigt.

**ANHANG C**

# *Abkürzungen*

| | |
|---|---|
| AR | Autoregressive Process |
| ARIMA | Autoregressive Integrated Moving-Average Process |
| ARMA | Autoregressive Moving-Average Process |
| BSC | Base Station Controler |
| BTS | Base Transceiver Station |
| CATI | Charging and Accounting Technology in the Internet |
| CHiPS | Connection-Holder-is-Preferred-Scheme |
| CLS | Conditional Least Squares |
| CORBA | Common Object Request Broker Architecture |
| CTMC | Continuous Time Markov Chain |
| DTMC | Discrete Time Markov Chain |
| EC | End Customer |
| EIR | Equipment Identity Register |
| ETSI | European Telecommunication Standard Institute |
| FARIMA | Fractal Autoregressive Integrated Moving-Average Process |
| FIFO | First In First Out |
| GMSC | Gateway Mobile Switching Center |
| GSF | Generalized Stitching Function |
| GSLO | Global Search Local Optimization |
| GSM | Global System for Mobile Communication |
| HLR | Home Location Register |
| IETF | Internet Engineering Task Force |
| i.i.d. | independent identically distributed |
| INDEX | Internet Demand Experiment |
| IP | Internet Protocol |

| | |
|---|---|
| IPP | Interrupted Poisson Process |
| ISP | Internet Service Provider |
| LA | Location Area |
| LU | Location Update |
| MA | Moving-Average |
| MAP | Markov Arrival Process |
| MMP | Markov Modulated Process |
| MMPP | Markov Modulated Poisson Process |
| MPEG | Moving Pictures Expert Group |
| MO | Mobile Originated Calls |
| MS | Mobile Station |
| MSC | Mobile Switching Center |
| MT | Mobile Terminated Calls |
| NBR | Nominal Bit Rate |
| OMG | Object Management Group |
| ORB | Object Request Broker |
| PCM | Pulse Code Modulation |
| PHB | Per Hop Behaviour |
| PMP | Paris Metro Pricing |
| PSP | Progressive Second Price |
| PSTN | Public Switched Telephone Network |
| QoS | Quality of Service |
| QoSP | Quality of Service Property |
| RFC | Request for Comments |
| RSVP | Resource Reservation Protocol |
| RUAA | Refined Uniform Asymptotic Approximation |
| SLA | Service Level Agreement |
| SMS | Short Message Service |
| SNF | Schweizer Nationalfond |
| SRDL | Service Request Description Language |
| SSF | Standard Stitching Function |
| SSP | State Setup Protocol |
| STM | Synchronous Transfer Mode |
| SUT | System under Test |
| TES | Transform-Expand-Sample |
| UAA | Uniform Asymptotic Approximation |
| ULS | Unconditional Least Squares |
| VLR | Visitor Location Register |
| VPN | Virtual Private Network |

**ANHANG D**

# *Abbildungs- und Tabellenverzeichnis*


Abb. 2-1 Elemente und Interfaces eines GSM-Netzwerks ..... 6
Abb. 2-2 Simulationsumgebung eines Mobile Switching Centers MSC als zu testendes System zwischen öffentlichem Netz und den Basisfunkstationen ..... 7
Abb. 2-3 Empirisch gemessener Verlauf von Location Updates über einen Zeitraum von fünf Werktagen ..... 8
Abb. 2-4 Varianten der empirischen Autokorrelationsfunktion und Randverteilungsdichte (Histogramm) der Referenzmessung ..... 10

Tab. 3-1 Zusammenhang zwischen empirischer Autokorrelationsfunktion und grenzstabilen Filtern .... 17
Abb. 3-2 Periodogramm der Referenzreihe ..... 26
Abb. 3-3 Partielle Autokorrelationen der Referenzreihe und Konfidenzintervall ..... 27
Tab. 3-4 Resultate der Schätzverfahren (a) – (d) für die ungefilterte Referenzreihe ..... 31
Abb. 3-5 AR(24)-Modell der Referenzreihe nach dem Yule-Walker-Verfahren und zugehörige Autokorrelationsfunktion ..... 31
Abb. 3-6 AR(24)-Modelle nach dem Householder- und Marquardt-Verfahren und zugehörige Autokorrelationsfunktionen ..... 32
Tab. 3-7 Box-Pierce Portmanteau-Statistik für die Beispielsmodelle ..... 33
Abb. 3-8 Erste Filterung der Referenzreihe, Autokorrelationen und Periodogramm ..... 34
Abb. 3-9 Zweite Filterung, Autokorrelationsfunktion und partielle Autokorrelationen ..... 34
Tab. 3-10 Parameterschätzung für das ARMA(2,2)-Modell der gefilterten Referenzreihe ..... 37
Abb. 3-11 Das ARMA(2,2)-Modell der Referenzreihe und zugehörige Autokorrelationsfunktionen ..... 37

Abb. 4-1 Das Schema des Standard-TES-Verfahrens ..... 41
Abb. 4-2 Einfachster Fall einer Innovationsdichte und Standard-Stitchingfunktionen ..... 42
Abb. 4-3 Exemplarische Innovationsdichte, Hintergrundsequenz, SSF und gestitchte Vordergrundsequenz ..... 43
Abb. 4-4 Zur Konvergenz von (4.3) in Abhängigkeit von der Anzahl der einbezogenen Summanden .... 44
Abb. 4-5 Zur Übereinstimmung von analytischer und empirischer Autokorrelation nach Melamed ..... 45
Abb. 4-6 Übereinstimmung von analytischer und empirischer Autokorrelation: eigenes Ergebnis für das gleiche Modell wie in Abbildung 4-5 ..... 46
Abb. 4-7 Driftlose Hintergrundsequenz (Unstetigkeiten vor dem Stitching und Standard-Stitching-Transformation mit Parameter 0.2 bzw. 0.8) ..... 47
Abb. 4-8 Einfache Innovationsverteilungen, driftende Hintergrundsequenzen und zugehörige Autokorrelationsfunktionen ..... 48

Abb. 4-9 Standard-TES-Modellierung der Referenzkurve und Autokorrelationsfunktion ..... 48
Abb. 4-10 Typisches Aussehen von Standard-TES-Modellen für die Referenzkurve und zugehörige Autokorrelationsfunktionen ..... 49
Abb. 4-11 Einfluß der Stitchingfunktion ..... 50
Abb. 4-12 Standard-Stitching-Funktion und zwei Kandidaten für eine Verallgemeinerte Stitching-Funktion ..... 51
Abb. 4-13 SSF und GSF und ihre Randverteilungen ..... 52
Abb. 4-14 Der Übergang von SSF zu GSF und seine Konsequenzen ..... 54
Abb. 4-15 GTES-Modell der Referenzkurve ..... 57
Abb. 4-16 GTES-Modell: Einfluß der Trägerbreite der Innovationsdichte ..... 57
Abb. 4-17 Progress-Retard-Modelle für die Referenzkurve ..... 58
Abb. 4-18 MPEG-Frames und ihre Abhängigkeit untereinander ..... 59
Abb. 4-19 Beispiel für eine Rahmensequenz am Beginn des Films “Das Schweigen der Lämmer” sowie zugehörige Autokorrelationsfunktion ..... 60
Abb. 4-20 Zwei Varianten einer GSF für das MPEG-Beispiel ..... 61
Abb. 4-21 Vordergrundsequenz des GTES-Modells für verschiedene Innovationsdichten und verschiedene GSF-Varianten sowie zugehörige Autokorrelationsfunktionen ..... 62
Abb. 4-22 Beide Varianten der GSF im Referenzbeispiel ..... 63
Abb. 4-23 Simulationsergebnisse für verschiedene δ und beide GSF-Varianten. ..... 64
Abb. 4-24 Modellierung der Referenzkurve mit dem automatisierten TES-Verfahren und zugehörige Autokorrelationsfunktionen ..... 69

Abb. 5-1 Auf dem Weg zu einer fraktalen selbstaffinen Kurve ..... 76
Abb. 5-2 Fraktales TES: Innovationsdichte, Hintergrundsequenz und gezoomte Vordergrundsequenzen 77
Abb. 5-3 Aggregate Variance-Plot für die Kurven aus Abbildung 5-2 ..... 77

Abb. 6-1 Dimensionen zur Klassifizierung von Preismodellen ..... 81
Abb. 6-2 Ein IntServ-over-DiffServ-Szenario ..... 90
Abb. 6-3 PATH-Message und RESV-Message entlang eines Verbindungspfads ..... 91
Abb. 6-4 Format einer RSVP-Nachricht ..... 92
Abb. 6-5 Format eines RSVP-Objekts ..... 92
Abb. 6-6 Anforderungen an ein Preismodell für Integrierte Internet-Dienste ..... 93

Abb. 7-1 Zur Strategie der Trunk Reservation ..... 97
Abb. 7-2 Beschreibung der Ressourcenauslastung als Geburts- und Todesprozeß ..... 98
Abb. 7-3 Die Funktion M maximal möglicher B-Gespräche ..... 98
Abb. 7-4 Preisfunktionen für zwei Gesprächsklassen ..... 100
Abb. 7-5 Preisfunktionen für C=120: Exakte Lösung, UAA und RUAA ..... 110
Abb. 7-6 Preisfunktionen für große Kapazitäten, eine stochastische Klasse und 90%, 95% bzw. 99% Last ..... 111
Abb. 7-7 Seiteneffekte in Szenarien mit einer stochastischen Klasse ..... 113
Abb. 7-8 Preisfunktionen für Mischklassen-Szenarien, jeweils UAA und RUAA ..... 114
Abb. 7-9 Preise für das Beispiel mit zwei stochastischen Klassen ..... 118
Abb. 7-10 Abhängigkeit der Preisfunktionen von Ankunftsraten und real vorliegender Kapazitätsaufteilung: RUAA für zwei stochastische Klassen ..... 119

Abb. 8-1 “Wer zu spät kommt, den belohnt das Leben”, jedenfalls bei Delta-Auktionen. Simulationsergebnisse und Vergleich der Zeiten für Protokolloverhead zwischen Delta-Auktionen und einem volumenbasierten Preisschema ..... 130
Abb. 8-2 Message Sequence Chart für CHiPS-Auktionen ..... 137
Abb. 8-3 Topologie des Simulationsszenarios ..... 140

Abb. 8-4 Marktpreisentwicklung im Fall einer Überlastung von Knoten 4 ..... 141
Abb. 8-5 Auslastung der Teilverbindung zwischen Knoten 0 und 4 bzw. Knoten 1 und 4. Nutzer 0 hat doppeltes Budget von Nutzer 1 zur Verfügung..... 142
Tab. 8-6 Vergleich Überlastfall mit und ohne CHiPS ..... 142
Abb. 8-7 Erhaltung der Incentive Compatibility unter CHiPS ..... 143

Abb. A-1 Zur Illustration der Inversionsmethode ..... 155
Abb. A-2 Allgemeine Stitching-Funktion ..... 156
Abb. A-3 Verteilungsfunktion der GSF aus Abbildung A-2 ..... 157

Abb. B-1 Das Prinzip des Tradings in einer CORBA-Umgebung ..... 171
Abb. B-2 Schnittstellenreferenzen angebotener und nachgefragter Dienste ..... 171
Abb. B-3 Assessment-Komponente und Evaluator-Komponente ..... 173
Abb. B-4 Mögliche Kurvenverläufe einer QoSP-Funktion ..... 175
Abb. B-5 Techniken zur Ermittlung der QoSP-Funktionen: (a) Direct Rating (b) Lineare Interpolation (c) Halbierungsmethode ..... 175
Abb. B-6 Laufzeitvergleich der zwei Versionen des erweiterten Traders ..... 177

**ANHANG E**

# *Literaturverzeichnis*


[AAB88] M. Ahamad, M. Ammer, J. Bernabeu-Auban, M. Khalidi: Using Multicast Communication to Locate Resources in a LAN-Based Distributed System. Proceedings of 13th Conference on Local Computer Networks, S. 193-202, Oktober 1988.

[APY98] H. Adiseshu, G. Parulkar, R. Yavatkar: A State Management Protocol for IntServ, DiffServ and Label Switching. ICNP 98, 6th IEEE International Conference on Network Protocols, October 1998.

[Bau91] H. Bauer: Wahrscheinlichkeitstheorie. De Gruyter Lehrbuch, 1991.

[BBC+98] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss: An Architecture for Differentiated Services. RFC 2475, IETF, Dezember 1998.

[BHJ48] E. Brockmeyer, H. L. Halstrom, A. Jensen: The Life and Works of A. K. Erlang. Academy of Technical Sciences, Kopenhagen, 1948.

[Ble66] N. Bleistein: Uniform Asymptotic Expansions of Integrals with Stationary Points near Algebraic Singularity. Commun. Pure Appl. Math., vol. 19, S. 353–370, 1966.

[BJ76] G. E. P. Box, G. M. Jenkins: Time Series Analysis: Forecasting and Control. Prentice-Hall 1976.

[BP70] G. E. P. Box, D. A. Pierce: Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. J. Amer. Statist. Ass. 65, 1509-1526, 1970.

[Bro97] N. Brownlee: Internet Pricing in Practice. In: [McKB97], S. 77-90.

[Bro98] D. Brocker: Messung und Modellierung komplexer Verkehrsstrukturen in Hochgeschwindigkeitsnetzen. Diplomarbeit, RWTH Aachen, September 1998.

[BS87] I. N. Bronstein, K. A. Semendjajew: Taschenbuch der Mathematik. Verlag Harri Deutsch, Thun und Frankfurt am Main 1987.

[BYF+99] Y. Bernet, R. Yavatkar, P. Ford, F. Baker, L. Zhang, M. Speer, R. Braden, B. Davie: Integrated Services Operation Over Diffserv Networks. Internet Draft draft-ietf-issll-diffserv-rsvp-02.txt, Juni 1999.

[BZB+97] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin: Resource Reservation Protocol (RSVP) - Version 1. Functional Specification. RFC 2205, IETF, Sept. 1997.

[CESZ93] R. Cocchi, D. Estrin, S. Shenker, L. Zhang: Pricing in Computer Networks: Motivation, Formulation and Example. IEEE/ACM Transactions on Networking, vol.1 no. 6, Dezember 1993, pp. 614-627.

[CF98] D. Clark, W. Fang: Explicit Allocation of Best-Effort Packet Delivery Service. IEEE/ACM Transactions on Networking, vol. 6 no. 4, August 1998.

[CL66] D. R. Cox, P. A. W. Lewis: The Statistical Analysis of Series of Events. Methuen & Co., London, 1966.

[Cop48] E. T. Copson: Theory of Functions of a Complex Variable. Oxford University Press 1948.

[corba] The Common Object Request Broker: Architecture and Specification. OMG Document 95.03.xx. Framingham (Mass.) 1995.

[Do97] M. Dochniak: Statistische Analyse von Leistungsmessungen in GSM-Netzen. Diplomarbeit Lehrstuhl Informatik 4, RWTH Aachen, 1997

[DS99] A. Dasylva, R. Srikant: Bounds on the Performance of Admission Control and Routing Policies for General Topology Networks with Multiple Call Classes. Proc. Infocom'99, New York, März 1999.

[Dur60] J. Durbin: The Fitting of Time-Series Models. Revue Inst. Int. de Statist., 1960.

[Ed99] R. Edell: The Internet Demand Experiment. Ph.D. thesis, University of California, Berkeley, Juni 1999.

[E-H97] M. Effer-Hack: Verkehrsmodellierung für GSM-Netze mit stochastischen Petrinetzen. Diplomarbeit, Lehrstuhl für Informatik 4, RWTH Aachen, 1997

[EV99] R. J. Edell, P. P. Varaiya: Providing Internet Access: What we learn from the INDEX Trial. Keynote Talk at Infocom '99 New York. INDEX Project Report #99-010W. URL: http://www. INDEX.Berkeley.EDU/99-010W.

[EW93] F. Eisenführ, M. Weber: Rationales Entscheiden. Berlin (Springer) 1993.

[Fal90] K. Falconer: Fractal Geometry. Mathematical Foundations and Applications. Wiley 1990.

[Fas98] A. Fasbender: Messung und Modellierung der Dienstgüte paketvermittelnder Netze. Dissertation RWTH Aachen, 1998.

[Flö98] F. Flössel: Classification of Internet Pricing Models. Semesterarbeit TIK, ETH Zürich, 1998.

[FM94] V. Frost, B. Melamed: Traffic Modeling for Telecommunications Networks. IEEE Communications Magazine (März 1994), 70–81.

[Fos99] M. Foser: Pricing- und Routen-Selektion für RSVP-Flows in einer Simulationsumgebung. Studienarbeit SA-99.03. TIK, ETH Zürich, 1999.

[FSP99] G. Fankhauser, D. Schweikert, B. Plattner: Service Level Agreement Trading for the Differentiated Services Architecture. Eingereicht zur Infocom 2000 Tel Aviv.

[FSVP98] G. Fankhauser, B. Stiller, C. Vögtli, B. Plattner: Reservation-based Charging in an Integrated Services Network. 4th INFORMS Telecommunications Conference, Bocca Raton, Florida, U.S.A., März 1998.

[GK95] R. J. Gibbens, F. P. Kelly: Network Programming Methods for Loss Networks. IEEE JSAC, vol. 13, no. 7, Sept. 1995, pp. 1189-1198.

[Gor99] M. Gorbatschow: Wie es war. Die deutsche Wiedervereinigung. München (Ullstein) 1999.

[GR93] R. J. Gibbens, P. Reichl: A General Performance Bound Applied to Examples of Highly Connected Loss Networks. Proc. of the 10th IEE UK Teletraffic Symposion, 3/1 - 3/11, April 1993.

[GR95] R. J. Gibbens, P. Reichl: Performance Bounds Applied to Loss Networks. In: D. M. Titterington (ed.): Complex Stochastic Systems and Engineering (S. 267-279). Oxford University Press 1995.

[Hak99] M. Hakim: Stochastische Modelle zur Beschreibung autokorrelierter Ankunftsströme in mobilen Telekommunikationsnetzen. Diplomarbeit RWTH Aachen, April 1999.

[Hav98] B. Haverkort: Performance of Computer Communication Systems: A Model-based Approach. Wiley 1998

[HEK98] J. Hartung, B. Elpelt, K.-H. Klösener: Lehr- und Handbuch der angewandten Statistik. Oldenbourg-Verlag 1998.

[HM99] S. van Hoesel, R. Müller: Optimization in Electronic Markets. Examples in Combinatorial Auctions. Internet Economy Workshop IEW'99. Berlin, Mai 1999.

[HMRS97] S. Hoff, P. Magnusson, P. Reichl, M. Schuba: Driving Performance Tests of Mobile Telecommunication Network Elements by Stochastic Traffic Models. Unveröffentlichtes Manuskript.

[Hui96] C. Huitema: IPv6: The New Internet Protocol. Prentice Hall, 1996.

[ILDK95] M. R. Ismail, I. E. Lambadaris, M. Devetsikiotis, A. R. Kaye: Modelling prioritized MPEG video using TES and a frame spreading strategy for transmission in ATM networks. Proceedings of the Infocom '95, 762 – 770.

[iso93] Coding of Moving Pictures and Associated Audio for Digital Storage Media up to 1.5 Mbit/s – Part 2: Video. International Standard: ISO/IEC IS 11172–2.

[JM92a] D. Jagerman, B. Melamed: The Transition and Autocorrelations Structure of TES Processes. Part I: General Theory. Commun. Statist. – Stochastic Models, 8(2), 193 – 219 (1992).

[JM92b] D. Jagerman, B. Melamed: The Transition and Autocorrelations Structure of TES Processes. Part II: Special Cases. Commun. Statist. – Stochastic Models, 8(3), 499 – 527 (1992).

[JM95] P. Jelenkovic, B. Melamed: Algorithmic Modeling of TES Processes. Rutcor Research Report 5-95, Feb. 1995.

[Kau81] J. S. Kaufman: Blocking in a shared resource environment. IEEE Trans. Comm., vol. COM-29, 1474-1481, 1981.

[Kel79] F. P. Kelly: Reversibility and Stochastic Networks. Wiley, 1979.

[Kel91a] F. P. Kelly: Effective bandwidths at multi-service queues. Queueing Systems 9 (1991) 5-16.

[Kel91b] F. P. Kelly: Loss Networks. Ann. Appl. Prob. 1, 319-378, 1991.

[Kel94] F. P. Kelly: Bounds on the Performance of Dynamic Routing Schemes for Highly Connected Networks. Mathematics of Operations Research, 19:1-20, 1994.

[Kel97] F. P. Kelly: Charging and rate control for elastic traffic. European Transactions on Telecommunications, vol. 8 (1997) 33-37.

[Key90] P. Key: Optimal Control and Trunk Reservation in Loss Networks. Probability in the Engineering and Informational Sciences, 4:203-242, 1990.

[Kil+98] K. Kilkki et al.: Internet Charging Reconsidered. 5th Annual Networld and Interop Engineers Conference. Las Vegas, Nevada, U.S.A., May 1998.

[Kle75] L. Kleinrock: Queueing Systems. Wiley 1975.

[KMT98] F. P. Kelly, A. K. Maulloo, D. K. H. Tan: Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research Society 49 (1998), 237-252.

[KO99] Y. A. Korilis, A. Orda: Incentive Compatible Pricing Strategies for QoS Routing. Proceedings Infocom ’99, New York, März 1999.

[KR76] R. Keeney, H. Raiffa: Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York (Wiley) 1976.

[KS98] F. P. Kelly, R. Steinberg: A combinatorial auction with multiple winners.

[KSWS98] M. Karsten, J. Schmitt, L. Wolf, R. Steinmetz: An Embedded Charging Approach for RSVP. International Workshop on QoS 98. Napa, California, U.S.A., Mai 1998.

[KZZ96] F. P. Kelly, S. Zachary, I. Ziedins: Stochastic Networks: Theory and Applications. Oxford University Press 1996.

[LCW97] D. Lam, D. C. Cox, J. Widom: Teletraffic Modeling for Personal Communications Services. IEEE Communications Magazine (Februar 1997), 79 – 87.

[LeG91] D. Le Gall: MPEG: A video compression standard for multimedia applications. Communications of the ACM 34 (4), 46–58.

[Lin98] C. Linnhoff-Popien: CORBA - Kommunikation und Management. Springer 1998.

[LM82] B. Liu, C. M. Munson: Generation of a Random Sequence Having a Jointly Specified Marginal Distribution and Autocovariance. IEEE Transactions on Acoustics, Speech and Signal Processing, no. 6, Dezember 1982.

[LMRS94] D. S. Lee, B. Melamed, A. R. Reibman, B. Sengupta: TES modeling for analysis of a video multiplexer. Performance Evaluation 16, 21–34.

[LRT97] C. Linnhoff-Popien, P. Reichl, D. Thißen: Including QoS Requirements into a CORBA Trader. ICODP/ICDP'97. Toronto, Mai 1997.

[LS97] A. A. Lazar; N. Semret: Auctions for Network Resource Sharing. CTR Tech. Report. Columbia University New York, Februar 1997.

[LS98] A. A. Lazar, N. Semret: Design, Analysis and Simulation of the Progressive Second Price Auction for Network Bandwidth Sharing. CTR Techn. Report. Columbia University New York, April 1998.

[LTWW94] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson: On the Self-Similar Nature of Ethernet Traffic (Extended Version). IEEE/ACM Transactions on Networking, vol. 2 no. 1 (1994) 1–15.

[Mar63] D. W. Marquardt: An Algorithm for Least-Sqares Estimation of Non-linear Parameters. In: Journal of the Society for Industrial and Applied Mathematics, 1963, S. 431.

[MB97] L. McKnight, J. Bailey (Hg.): Internet Economics. MIT Press, Cambridge (Massachusetts), 1997.

[Mel91] B. Melamed: TES: A Class of Methods for Generating Autocorrelated Uniform Variates. ORSA Journal on Computing 3 (4), 1991.

[Mel93] B. Melamed: An Overview of TES Processes and Modeling Methodology. In: Donatiello, L.; Nelson, R. (eds.): Performance Evaluation of Computer and Communications Systems. Springer (LNCS) 1993, 359–393.

[MGH91] D. Mitra, R. J. Gibbens, B. D. Huang: Analysis and Optimal Design of Aggregated-Least-Busy-Alternative Routing on Symmetric Loss Networks with Trunk Reservations. In: A. Jensen, V. B. Iversen (Hg.): Teletraffic and Datatraffic in a Period of Change, ITC-13, S. 477-482. Elsevier 1991.

[MLB93] Z. Milosevic, A. Lister, M. Bearman: New economic-driven aspects of the ODP Enterprise specification and related Quality of Service issues. In: J. de Meer, B. Mahr, S. Storp (Hg..): Open Distributed Processing II. Proceedings of the IFIP TC6/WG6.1 International Conference on Open Distributed Processing. Berlin, September 1993 (Chapman & Hall).

[MM87] R. Preston McAfee, H. McMillan: Auction and Bidding. Journal of Economic Literature, 25: 699-738, 1987.

[MM94] D. Mitra, J. A. Morrison: Erlang Capacity and Uniform Approximations for Shared Unbuffered Resources. IEEE/ACM Transactions on Networking vol. 2 no. 6, Dezember 1994.

[M-M97] J. MacKie-Mason: A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network. University of Michigan, September 1997.

[MMM95] J. MacKie-Mason, J. Murphy, L. Murphy: The Role of Responsive Pricing in the Internet. Tech. Report, University of Michigan, June 95. In: [MB97].

[MMR96] D. Mitra, J. Morrison, K. Ramakrishnan: ATM Network Design and Optimization: A Multirate Loss Network Framework. IEEE/ACM Transactions on Networking, vol.4 no. 4, August 1996.

[MMR99] D. Mitra, J. A. Morrison and K. G.Ramakrishnan: Optimization and Design of Network Routing using Refined Asymptotic Approximations. Proc. Performance 99.

[Moh99] F. Mohren: Verkehrsmodellierung mit TES (Transform-Expand-Sample). Diplomarbeit, Lehrstuhl für Informatik 4, RWTH Aachen, Febr. 1999.

[MP94] M. Mouly, M. B. Pautet: The GSM System for Mobile Communication. Veröffentlicht im Eigenverlag, 4 rue Elisée Reclus, F-91120 Palaiseau, France, 1994.

[MRM98] J. Morrison, K.G. Ramakrishnan, D. Mitra: Refined asymptotic approximations to loss probabilities and their sensitivities in shared unbuffered resources. SIAM Journal Appl. Math., März 1998.

[MRS99] F. Mohren, P. Reichl, M. Schuba: Automatisierte Verkehrsmodellierung mit TES. Unveröffentlichtes Manuskript. Lehrstuhl Informatik 4, RWTH Aachen, 1999.

[MS93] B. Melamed, B. Sengupta: TES Modeling of Video Traffic. Technical Report. NEC USA Inc., C&C Research Laboratories, Princeton (NJ), Dezember 1993.

[MV93] J. MacKie-Mason, H. Varian: Pricing the Internet. JFK School for Government, Mai 1993.

[MZP96] B. Meyer, S. Zlatintsis, C. Popien: Enabling Interworking between Heterogeneous Distributed Platforms. Proceedings of ICDP'96. Dresden, Februar 1996 (Chapman & Hall).

[Neb81] O. Nebel: Das dichterische Werk. Edition Text und Kritik München 1981.

[ns-2] UCB/LBNL/VINT Network Simulator ns (version 2). http://www-mash.cs.berkeley.edu/ns.

[Odl97] A. M. Odlyzko: A modest proposal for preventing Internet congestion. URL: http://www.research.att.com/~amo/doc/complete.html

[Odl99] A. M. Odlyzko: Paris Metro Pricing for the Internet. Proc. ACM Conference on Electronic Commerce (EC-99), ACM, 1999.

[omg95] ISO/IEC JTC1/SC21 WG7 (ODP): ODP Trader Document. OMG Document Number 95-07-06, 1995.

[orbix96] IONA Technologies Ltd.: Orbix - Programmer's Guide and Reference Manual. Release 2.0, 1996.

[PM95] C. Popien, B. Meyer: A Formal Approach to Service Import in ODP Trader Federations. In: D. Hogrefe, S. Leuer: Formal Description Techniques VII. Chapman & Hall 1995.

[Po95] C. Popien: Dienstvermittlung in Verteilten Systemen. Teubner Texte zur Informatik. Stuttgart (Teubner) 1995.

[PSW96] C. Popien, G. Schürmann, K.-H. Weiß: Verteilte Verarbeitung in Offenen Systemen. Stuttgart (Teubner) 1996.

[qos95] ISO/IEC JTC1/SC21/N9309: Open Systems Interconnection, Data Management and Open Distributed Processing – Quality of Service, Basic Framework. Working Draft, Januar 1995.

[Rei92] P. Reichl: Loss Networks. A General Bound for their Performance Investigated by Some Simulations. Part III Essay, University of Cambridge (UK), 1992.

[Rei94] P. Reichl: Eine allgemeine untere Schranke für die Verlustrate in nicht-symmetrischen Netzwerken. Diplomarbeit TU München, Mai 1994.

[Rei98a] P. Reichl: Does the TES Stitching Function Merely Stitch? Proceedings of the IPCCC'98. Phoenix, AZ, Feb. 1998.

[Rei98b] P. Reichl: A GTES Model for Periodic Traffic. Proceedings of the ICC'98. Atlanta, GA, Juni 1998.

[Rei99a] P. Reichl: Approximated Price Functions for Dynamic Volume-based Pricing of Multiclass Internet Traffic. Proceedings MMB'99 Trier, Sept. 1999.

[Rei99b] P. Reichl: Kelly's Bound, RUAA and the Pricing of Multiclass Traffic in Loss Networks. Proceedings of UK Performance Engineering Workshop '99, Bristol, Juli 1999.

[RFS99] P. Reichl, G. Fankhauser, B. Stiller: Auction Models for Multiprovider Internet Connections. Proceedings MMB'99 Trier, Sept. 1999.

[RKJS98] P. Reichl, D. Kesdogan, K. Junghärtchen, M. Schuba: Simulative Performance Evaluation of the Temporary Pseudonym Method for Protecting Location Information in GSM Networks. Proceedings of TOOLS'98, Palma de Mallorca, Sept. 1998.

[RLeB97] S. Robert, J.-Y. LeBoudec: New Models for Pseudo Self-Similar Traffic. Performance Evaluation, 30, S. 57-68, 1997.

[RLS99] P. Reichl, S. Leinen, B. Stiller: A Practical Review of Pricing and Cost Recovery for Internet Services. Internet Economy Workshop IEW'99, Berlin, Mai 1999.

[RLT97] P. Reichl, C. Linnhoff-Popien, D. Thißen: Einbeziehung von Nutzerinteressen bei der QoS-basierten Dienstvermittlung unter CORBA. Proc. of KiVS'97, Braunschweig 1997.

[RMR94] D. Reininger, B. Melamed, D. Raychaudhuri: Variable bit rate MPEG video: Characteristics, modeling and multiplexing. Proceedings of ITC-14, 295 – 306.

[Ro95] O. Rose: Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. University of Würzburg. Institute of Computer Science Research Report Series. Report No. 101. Februar 1995.
Die MPEG-Daten sind unter ftp-info3.informatik.uni-wuerzburg.de/pub/MPEG/ verfügbar.

[Ro96] O. Rose: Estimation of the Hurst Parameter of Long-Range Dependent Time Series. Research Report No. 137, University of Würzburg, Institute of Computer Science, Feb. 1996.

[Ro97] O. Rose: Traffic Modeling of Variable Bit Rate MPEG Video and its Impacts on ATM Networks. Ph.D. thesis. Würzburger Beiträge zur Leistungsbewertung Verteilter Systeme. Bericht 02/97. Würzburg 1997.

[Rob82] J. W. Roberts: Teletraffic models for the Telecom 1 integrated services network. Proceedings of the 10th International Teletraffic Congress 1982, ITC-10, session 1, paper # 2.

[Rog98] R. Rogerson: Usage-related Charges for the JANET Network. JISC Circular 3/98, März 1998. URL: http://www.jisc.ac.uk/pub98/c3_98.html.

[Ros95] K. W. Ross: Multirate Loss Models for Broadband Telecommunications Networks. Springer, New York, 1995.

[RPH97] M. H. Rothkopf, A. Pekec, R. M. Harstad: Computationally manageable combinational auctions. Management Science, 1997.

[RSB82] S. J. Rassenti, V. L. Smith, R. L. Bulfin: A combinatorial auction mechanism for airport time slot allocation. Bell Journal of Economics 13 (1982) 402-417.

[RSH97] P. Reichl, S. Hoff, M. Schuba: How to Model Complex Periodic Traffic with TES. Proceedings of the 13th UK Workshop on Performance Engineering, Ilkley, West Yorkshire, July 1997. Edinburgh University Press 1997, S. 17/1 - 17/11.

[RSL99] P. Reichl, B. Stiller, S. Leinen: Pricing Models for Internet Traffic. Technical Report. CATI-TIKSWI-DN-P-008_2.2, TIK, ETH Zürich, April 1999.

[RTL96] P. Reichl, D. Thißen, C. Linnhoff-Popien: How to Enhance Service Selection in Distributed Systems. Proc. of the International Conference on Distributed Computer Communication Networks DCCN‘96, 114 - 123. Tel Aviv, Israel, Nov. 1996.

[SBC94] S. Shenker, R. Braden, D. Clark: Integrated Services in the Internet Architecture: An Overview. RFC 1633, Juni 1994

[SBGP99] B. Stiller, T. Braun, M. Günter, B. Plattner: The CATI Project – Charging and Accounting Technology for the Internet. 4th European Conference on Multimedia Applications, Services, and Techniques (ECMAST’99), Madrid, LNCS Vol. 1629, Springer, Berlin, Mai 1999, 281-296.

[Sch93] R. Schlittgen: Einführung in die Statistik: Analyse und Modellierung von Daten. Oldenbourg-Verlag 1993.

[SCEH96] S. Shenker, D. Clark, D. Estrin, S. Herzog: Pricing in Computer Networks: Reshaping the Research Agenda. ACM Computer Communications Review, Vol. 26, No. 2, April 1996 S. 19-43.

[SchS89] R. Schlittgen, B. H. Streitberg: Zeitreihenanalyse. R. Oldenbourg-Verlag, 1989.

[Schw99] D. Schweikert: QoS Routing and Pricing in Large Scale Internetworks. Diplomarbeit TIK, ETH Zürich, März 1999.

[SFJ+99] B. Stiller, G. Fankhauser, G. Joller, P. Reichl, N. Weiler: Open Charging and QoS Interfaces for IP Telephony. INET‘99, San Jose (CA), Juni 1999.

[SFPN98] B. Stiller, G. Fankhauser, B. Plattner, N. Weiler: Charging and Accounting for Integrated Internet Services - State of the Art, Problems and Trends. Proceedings of INET’98, Juli 1998.

[She95] S. Shenker: Some Fundamental Design Issues for the Future Internet. IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, September 1995 pp. 1176-1188.

[Sp99] B. Spielmann: Design and Evaluation of Efficient Pricing Schemes for Integrated Internet Services. Diplomarbeit ETH Zürich. Juli 1999.

[SPM94] O. Spaniol, C. Popien, B. Meyer: Dienste und Dienstvermittlung in Client/-Server-Systemen. Thomsons Aktuelle Tutorien. International Thomson Publishing 1994.

[SPRS96] O. Spaniol, C. Popien, P. Reichl, M. Schuba: Systemprogrammierung. Augustinus-Verlag Aachen, 1996.

[SRL99] B. Stiller, P. Reichl, S. Leinen: Pricing and Cost Recovery for Internet Services: Practical Review, Classification, and Application of Relevant Models. NETNOMICS'99.

[SSRS98] O. Spaniol, M. Schuba, P. Reichl, G. Schneider. Lokale Netze. Augustinus-Verlag Aachen, 1998.

[Sto93] J. Stoer: Numerische Mathematik 1 und 2. Springer 1993.

[Tan96] A. S. Tanenbaum: Computer Networks. 3rd Edition. Prentice Hall 1996.

[Thi96] D. Thißen: QoS-basierte Optimierung der Dienstselektion in einem Orbix-Trader. Diplomarbeit Lehrstuhl Informatik IV der RWTH Aachen, Juli 1996.

[TP96] D. Thißen, C. Linnhoff-Popien: Finding Optimal Services within a CORBA Trader. In: Trends in Distributed Systems. CORBA and Beyond. Aachen, Oktober 1996 (Springer).

[Tra98] T. Trajkovska: Modellierung von zyklischem und selbstähnlichem Netzverkehr mit Hilfe von autoregressiven Prozessen. Diplomarbeit Lehrstuhl für Informatik 4, RWTH Aachen, 1998.

[TTW95] M. S. Taqqu, V. Teverovsky, W. Willinger: Estimators for long-range dependence: an empirical study. Fractals Vol. 3 No. 4 (1995), 785 – 788.

[Vick61] W. Vickrey: Counterspeculation, auctions and competitive sealed tenders. Journal of Finance, 1961.

[Vög98] C. Vögtli: Auktions-basierte Reservationen und Verrechnung für das Next Generation Internet. Diplomarbeit ETH Zürich, März 1998.

[WTE96] W. Willinger, M. Taqqu, A. Erramilli: A Bibliographic Guide to Self-Similar Traffic and Performance Modelling for Modern High-Speed Networks. In: [KZZ96] 339-366.

[WWWM98] M. Wellman, W. Walsh, P. Wurman, J. MacKie-Mason: Auction Protocols for Decentralized Scheduling. Extended Version of "Some economics of market-based distributed scheduling", vorgestellt bei der 18th International Conference on Distributed Computing Systems, Amsterdam, Mai 1998.

[Yul27] G. U. Yule: On a Method of Investigating Periodicities in Disturbed Series with Special Reference to Wolfer's Sunspot Numbers. Phil. Trans., 1927.

[ZDE+93] L. Zhang, S. Deering, D. Estrin, S. Shenker, D. Zappala: RSVP: A New Resource ReServation Protocol. IEEE Networks Magazine vol. 31 no. 9, Sept. 1993, 8-18.

# Lebenslauf

| | |
|---|---|
| 14. Mai 1967 | geboren in Pappenheim (Bayern) |
| 1986 | Abitur am Werner-von-Siemens-Gymnasium Weißenburg (Bay.) |
| 1987 – 1994 | Studium von Mathematik, Physik und Philosophie an der Technischen Universität bzw. der Hochschule für Philosophie München |
| 1989 | Vordiplom in Mathematik |
| 1990 | Vordiplom in Physik |
| 1991 | Bakkalaureat in Philosophie |
| 1991 – 1992 | Masters-Studium an der University of Cambridge (Großbritannien) Certificate of Advanced Studies in Mathematics |
| 1992/93 | wiederholte mehrmonatige Forschungsaufenthalte am Statistical Laboratory der Universität Cambridge |
| 1994 | Abschluß als Diplom-Mathematiker (Univ.) an der TU München |
| 1995 – 1998 | Wissenschaftlicher Mitarbeiter am Lehrstuhl für Informatik IV (Kommunikationssysteme) der RWTH Aachen |
| 1996/97 | Lehraufträge “Netzmanagement” und “Verteilte Systeme” im Studiengang Wirtschaftsinformatik der Universität GH Essen |
| 1998 | mehrmonatiger Forschungsaufenthalt bei Bell Labs (Lucent Technologies) in Murray Hill (New Jersey) |
| 1998 – 2000 | Wissenschaftlicher Gast am Institut für Technische Informatik und Kommunikationssysteme (TIK) der ETH Zürich |