Thema:

Entwicklung eines Forecastsystems für Call Center mit Best Service Routing

Diplomarbeit

Arbeitsgruppe Wirtschaftsinformatik

Themensteller: Prof. Dr. rer. pol. habil. Hans-Knud Arndt

Betreuer: Dipl.-Kfm. Henner Graubitz

Vorgelegt von: Rick Heiderich

Abgabetermin: 03. September 2008

Inhaltsverzeichnis

Inhaltsverzeichnis II
Verzeichnis der Abkürzungen und Akronyme IV
Symbolverzeichnis V
Abbildungsverzeichnis VI
Tabellenverzeichnis VII
1 Kurzfassung 1
2 Motivation 2
3 Call Center Grundlagen 4
3.1 Begriffsbildung 4
3.1.1 Was ist ein Call Center? 4
3.1.2 Call Center Agents 4
3.1.3 Inbound 5
3.1.4 Outbound 6
3.2 Wachstum als Leitmotiv der Call Center Branche 6
3.3 Die Personaleinsatzplanung 8
3.3.1 Einflußgrößen der Einsatzplanung 8
3.3.2 Der Servicelevel 9
3.3.3 Workforce-Management-Systeme 10
3.3.4 Prognoseanforderungen beim Best Service Routing 12
4 Die Warteschlangentheorie 14
4.1 Begriffsabgrenzung 14
4.2 Littles Gesetz 15
4.3 Die Kendall-Notation 15
4.4 Markov-Prozesse 16
4.5 Die praktische Anwendung der Warteschlangentheorie 18
5 Zeitreihenanalyse 20
5.1 Zeitreihen und stochastischen Prozessen 20
5.2 Überblick zu Ansätzen der Prognosebildung 20
5.3 Die exponentielle Glättung zweiter Ordnung 21
5.4 Konfidenzintervalle 22
6 Die Personalbedarfsprognose im Call Center 24
6.1 Ausgangsdaten der Personalbedarfsprognose 24
6.2 Die Prognose des Anrufaufkommens 26
6.3 Die Produktivität der Agents 27
6.4 Bedarfsrechnung mit Erlang-C 28
6.5 Eine Datenanalyse 31
6.6 Erweiterung des Erlang-C Modells 33

6.7 Ein Optimierungsmodell für Vertriebs-Call-Center 40
6.8 Probleme der Bedarfszahlen 42
7 Best Service Routing 44
7.1 Routingvarianten 44
7.2 Das Modell des Best Service Routings 46
7.3 Fragestellungen 47
8 Simulation des Best Service Routing 49
8.1 Wahl des Simulationsart 49
8.1.1 Vorüberlegungen zum Simulationsmodell 50
8.1.2 Objekte und Variablen des Simulationsmodells 50
8.2 Die Inversionsmethode 54
8.3 Der Prozessablauf der Simulation 56
8.4 Ergebnisse der Simulation 58
9 Das finale Optimierungsmodell des BSR 62
10 Ausblick 64
Literaturverzeichnis 67

Verzeichnis der Abkürzungen und Akronyme

ACD	Automatic Call Distribution
API	Application Programming Interface
BAP	Bildschirmarbeitspause
BPO	Business Process Outsourcing
BSR	Best Service Routing
CC	Call Center
Db	Deckungsbeitrag
DBMS	Datenbankmanagementsystem
GUI	Graphical User Interface
HTML	Hypertext Markup Language
IB	Inbound
IDE	Integrated Development Environment
ISO	International Organisation of Standardization
JAZ	Jahresarbeitszeit
KNN	künstliches neuronales Netz
LKC	Leiter Kundencenter
MB	Megabyte
MC	Markov-Chain
mGz	mittlere Gesprächszeit
mKz	mittlere Klingelzeit
mSz	mittlere Sperrzeit
MS	Microsoft
NBI	Neutraler Bericht im Intervall
OB	Outbound
ODBC	Open Database Connectivity
PL	Projektleiter
RDBMS	Relationales Datenbankmanagementsystem
SB	Sachbearbeiter
Sek	Sekunden
SQL	Structured Query Language
TL	Teamleiter
UML	Unified Modelling Language
VBA	Visual Basic for Applications
VPN	Virtual Privat Network

Symbolverzeichnis

AnkAnr	Ankommende Anrufe (im Intervall)
AUC	durchschnittlicher Umsatz pro Anruf
AWT	durchschnittliche Wartezeit
AHT	durchschnittliche Anrufdauer
Bes	Besetztfälle (im Intervall)
c	Agentanzahl
e	Eulersche Zahl (e = 2,718281828459…)
E	Zustandsmenge einer Markov-Chain
L	Lohn eines Agenten pro Zeiteinheit
mGz	mittlere Gesprächszeit
mKz	mittlere Klingelzeit
mSz	mittlere Sperrzeit
P	Übergangsmatrix einer Markov-Chain
Prod	Produktivität
S	Index des Standortes
T	Intervalllänge einer Planungszeiteinheit
v	Auflegerrate
Verz	Verzichter
μ	Bedienrate
λ	Ankunftsrate der Anrufe
σ	Standardabweichung

Abbildungsverzeichnis

Abb. 3.1: Komponenten von Workforce-Management-Systemen 10
Abb. 4.1: Ein einfaches Wartesystem 14
Abb. 4.2: Geburts- und Sterbeprozess 18
Abb. 4.6.1: Wochenverlauf des Anrufaufkommens 26
Abb. 6.2: Die Erlang-C Formel 29
Abb. 6.3: Dichtefunktion der Wartezeitverteilung 30
Abb. 6.4: Wartezeiten in der Queue 35
Abb. 6.5: gemeinsame Verteilungsfunktion 37
Abb. 6.6: Aufteilung der Anrufergebnisse 39
Abb. 7.1: Modell des Best Service Routings 46
Abb. 8.1: Visualisierung der Simulation 53
Abb. 8.2: Inversionsmethode für Exponentialfunktionen 55
Abb. 8.3: Prozess eines eintreffenden Anrufs 57
Abb. 8.4: Veränderungen des Servicelvels 58
Abb. 8.5: Umsatzkurven 59

Tabellenverzeichnis

Tab. 8.1: Ergebnisreihen der BSR Simulation 60

1 Kurzfassung

Die vorliegende Arbeit beschäftigt sich mit einem speziellen Problem der Personalbedarfsprognose von Call Centern. Im Gegensatz zum Großteil der in der Fachliteratur anzutreffenden Ansätzen, bei denen meist ein System mit nur einem Call Center betrachtet wird, soll hier ein mit dem Begriff Best Service Routing bezeichnetes System analysiert werden. Dabei stehen mehrere Standorte in vernetzter Beziehung und wirken wechselseitigen Einfluss bezüglich der jeweils erhaltenen Anrufmengen aufeinander aus. Beim reinen Best Service Routing bekommt immer derjenige Standort den eintreffenden Anruf, der zum aktuellen Zeitpunkt die besten Voraussetzungen für dessen schnellstmögliche Bearbeitung bietet.

Die Erstellung einer Personalbedarfsprognose für ein einzelnes Call Center dieses Teilnehmerkreises ist schwierig, da die Teilnehmer im Normalfall Profitcenter sind und als solche unabhängig voneinander agieren. Es muss also davon ausgegangen werden, dass zwar über den eigenen Standort genügend Informationen bereitstehen, aber Informationen über andere Standorte zum Großteil nur durch Schätzung mittels spezieller mathematischer Verfahren zugänglich sind.

Ausgehend von der Klärung grundlegender Begrifflichkeiten und Sachverhalten der Call Center Branche wird der Leser mit wichtigen Modellen der Wahrscheinlichkeits- und Warteschlangentheorie, gängigen Prognosemethoden und Verfahren der Personalbedarfsrechnung bekannt gemacht. Im Anschluss an die Erweiterung eines bekannten Verfahrens der Personalbedarfsermittlung soll im Rahmen einer Simulation tiefer auf das eigentliche Problem des Best Service Routings eingegangen werden. Die dabei erzielten Ergebnisse liefern die Grundlage für die Konkretisierung eines Optimierungsmodells und eines Prognoseverfahrens für die Bestimmung eines möglichst optimalen Mitarbeiterbedarfs im Falle des Best Service Routings.

2 Motivation

Der Bedarf nach Prognosen und Prognosemethoden existiert, seit Menschen voraus denken um zukünftige Handelungen zu planen. In antiken Zeiten befragten Herrscher ihre Seher und Orakel, um Ereignisse vorherzusehen. Man erhoffte sich Aussagen über das kommende Wetter bis hin zum Ausgang entscheidender Schlachten. Das Vorgehen der frühen Prognosebeauftragten hatte, abgesehen von der psychologischen Beeinflussung, nur wenig mit heutigen wissenschaftlichen Methoden gemeinsam. Meist handelte es sich um Geisterbeschwörungen und einfache Zufallsspiele, bei denen beispielsweise Knochen geworfen wurden, um aus ihrer Anordnung die Zukunft zu lesen.

Der Glaube an göttliche Einblicke ist heute größtenteils der Erkenntnis gewichen, dass viele Ereignisse im Voraus nicht genau bestimmt werden können. In der Physik beispielsweise stellt die Unbestimmtheit auf Quantenebene eine Grenze des menschlichen Wissens dar. Aber dort, wo Beobachtungen nicht erklärt werden können und als zufällig gelten, liefern zahlreiche Entwicklungen, auf den Gebieten Wahrscheinlichkeitstheorie und Statistik, Methoden, um mit der Unbestimmtheit zurechtzukommen. Mit Hilfe so genannter Markov-Ketten werden stochastische Prozesse beschrieben und bedingte Wahrscheinlichkeiten und Verteilungsfunktionen liefern das Handwerkszeug, um mit dem Zufall zu rechnen. Bei hoher Komplexität kann eine umfassende mathematische Beschreibung allerdings häufig nicht mehr realisiert werden. In diesem Fall bieten Simulationsverfahren einen Analyseansatz zur näherungsweisen Bestimmung und Vorhersage eines Systemverhaltens. Die statistische Auswertung vieler Beobachtungen lässt darüber hinaus Vorhersagen zu, die den realen Ergebnissen zumindest im Durchschnitt näher kommen als intuitives Raten.

Auch in der heutigen Wirtschaft spielen Prognosen eine überaus bedeutende Rolle. Bei Aktienanalysen werden modernste Prognoseverfahren angewendet um Kursentwicklungen vorherzusehen und auch innerhalb von Unternehmen müssen beispielsweise Vorhersagen zum Umsatz und häufig auch strategische Entscheidungen unter zum Teil erheblicher Unsicherheit getroffen werden. Die Basis derartiger Entscheidungen sind häufig Prognosen zur Entwicklung möglicher Szenarien. Wer hierbei genauere Vorhersagen trifft und Risiken angemessen einkalkuliert, kann wirtschaftlicher operieren und sich so gegen andere Wettbewerber behaupten.

Die Grundvoraussetzung für eine Prognose ist immer eine möglichst genaue Kenntnis der jeweils wirkenden Einflussfaktoren und deren Abhängigkeiten untereinander. Ein eingesetztes Prognoseinstrument sollte also auf einem Modell aufbauen, das die Umgebung des Problems hinreichend genau abbildet. Es muss flexibel auf mögliche

Umweltzustände reagieren können und sollte transparent und hinsichtlich seiner Richtigkeit und Genauigkeit überprüfbar sein. Für ein Unternehmen ist das entscheidende Kriterium aber letztendlich immer der betriebswirtschaftliche Mehrwert, den beispielsweise die Einführung einer neuen Prognosesoftware verursacht. Da dieser im Allgemeinen relativ schwer messbar ist, sollten zumindest die aus der erweiterten Automatisierung und einer hoffentlich verbesserten Prognosegenauigkeit resultierenden Vorteile den Entwicklungsaufwand rechtfertigen.

Im Falle des Best Service Routings gibt es zahlreiche Unsicherheiten, die beispielsweise die Bestimmung eines optimalen Mitarbeitereinsatzes erschweren. Obwohl das Anrufaufkommen des Gesamtsystems mit herkömmlichen statistischen Verfahren sehr gut prognostizierbar ist, kann es aufgrund der fehlenden Abstimmung der teilnehmenden Call Center sehr leicht vorkommen, dass es in einzelnen Tagesbereichen zu starker Über- oder Unterbesetzung kommt. Das heißt, dass bezogen auf die Gesamtbesetzung aller beteiligten Call Center entweder zu viele oder zu wenige Mitarbeiter zur Verfügung stehen. Im Fall von zu viel eingesetzten Mitarbeitern sitzen einige untätig herum und es entstehen unnötige Kosten für die Betreiber der Call Centers. Man spricht in diesem Zusammenhang auch von Bereitzeiten. Im umgekehrten Fall, mit zu wenigen Mitarbeitern, bedeutet dies ein Ärgernis für die anrufenden Kunden, da diese länger warten müssen, bis sie bedient werden. Gerade die Steigerung der Servicequalität ist aber ein Ziel, das sich die Call Center Betreiber durch die Einführung von Best Service Routing erhoffen. Die Grundidee des Systems ist eine ausbalancierte Gleichbelastung aller Standorte, um so die Engpässe, die oft in kleineren Call Centern auftreten, abzufangen. Theoretisch ist diese Hoffnung begründet, die Praxis zeigt aber leider auch viele Schwierigkeiten. Probleme können vor allem durch die Eigenverantwortlichkeit der einzelnen Call Center für die zahlenmäßige Ausstattung und zeitliche Verplanung ihres Personals auftreten.

Die Aufgabe ist also, eine Methode zu entwickeln, die dem einzelnen Call Center, trotz geringer Abstimmung mit den anderen Teilnehmern des Best Service Routings, eine möglichst gute Personalbedarfsprognose ermöglicht.

3 Call Center Grundlagen

3.1 Begriffsbildung

3.1.1 Was ist ein Call Center?

Bevor tiefer auf das eigentliche Modell eingegangen wird, das dem neuen Prognosesystem zugrunde liegt, sollen an dieser Stelle einige Begrifflichkeiten erläutert werden, die für ein einheitliches Verständnis der Call Center Branche von Bedeutung sind.

Call Center sind Dienstleistungsbetriebe, in denen sogenannte Tele-Dienste produziert werden. Diese Dienste sind gekennzeichnet durch eine räumliche Trennung bei gleichzeitig zeitlicher Aneinanderbindung von Konsument und Produzent des Dienstes. (vgl. Helbert (2003), S. 3)

Werden neben den Telefondiensten auch Dienstleistungen über weitere Medien wie Email oder Fax angeboten, so spricht man heute überwiegend von einem Contact Center.

Unterschiedliche Arten von Call Centern lassen sich vor allem aus den unterschiedlichen Aufgaben ableiten, für die sie eingesetzt werden. Neben der klassischen Telefonauskunft, die ein sogenanntes Dienstleistungs-Call-Center darstellt, existieren als weitere Formen unter anderem noch das Vertriebs-Call-Center und das Betreuungs-Call-Center. Das Vertriebs Call Center ist schon seinem Namen nach für den Vertrieb kommerzieller Erzeugnisse vorgesehen. Beispiele dafür sind die telefonische Bestellannahme eines Versandhandels oder der aktive Vertrieb von beispielsweise Mobilfunk- oder Festnetzverträgen. Betreuungs-Call-Center sind vor allem für den Kundensupport zuständig und könnte zum Beispiel für die Betreuung einer kostenlosen Hotline zuständig sein.

3.1.2 Call Center Agents

Die Mitarbeiter in der Produktion eines Call Centers, welche den eigentlichen Telefondienst leisten, werden als Agents oder Kundenberater bezeichnet. Die Agents erhalten durch Schulungen bestimmte Befähigungen, auch Skills genannt, die sie für die Ausübung unterschiedlicher Arten von Tele-Diensten qualifizieren. Für den Einsatz von Best Service Routing geht diese Arbeit von der einfachsten Variante aus, bei der alle Agents dieselbe Art von Anrufen bearbeiten, weshalb keine Unterschiede in den Qualifikationen berücksichtigt werden müssen.

Ein Agent arbeitet im Normalfall an einem rechnergestützten Arbeitsplatz, von denen es in einem Call Center mehrere hundert Stück geben kann. Alle Arbeitsplätze sind dabei über ein betriebsweites Netzwerk verbunden.

Neuerdings existieren Lösungen, die den menschlichen Agent zumindest in einigen Bereichen ersetzen können. Derartige Systeme sind unter dem Namen Voice Portale und Interactive Voice Response (IVR) beispielsweise im Bereich verschiedener telefonischer Auskünfte im Einsatz und basieren auf Technologien zur Spracherkennung. Im Rahmen der Personalbedarfsprognose wird in dieser Arbeit aber immer von menschlichen Agents ausgegangen. Gerade der menschliche Faktor erhöht die Komplexität des Problems und muss sowohl auf der Seite der Anrufer, als auch auf der Seite der Agents berücksichtigt werden.

3.1.3 Inbound

Allgemein wird beim Call Center Betrieb zwischen den zwei grundlegenden Formen Inbound und Outbound unterschieden. Im Inbound (IB) kommen Anrufe, auch Calls genannt, von außen herein. Ein anschauliches Beispiel dafür ist wieder die Telefonauskunft, bei der Privatpersonen und Geschäftskunden anrufen, um Telefonnummern zu erfragen.

Eine ACD-Anlage übernimmt im IB-Betrieb das interne Routing der Calls, also das Verteilen eintreffender Anrufe auf die Agents. ACD steht für Automatic Call Distribution, also für eine automatische Anrufverteilung. Ein ankommender Anruf wird im Idealfall an einen freien Agent mit dem entsprechenden Skill für die Bearbeitung weitergeleitet.

Steht bei Ankunft eines Anrufs kein freier Agent zur Verfügung, so wird dieser Anruf in eine Warteschlange, die sogenannte Queue, eingereiht. Dort verbleibt er, bis er an der Reihe ist, angenommen zu werden, oder bis der Anrufer aus Ungeduld vorzeitig auflegt. Die Queue selbst hat eine endliche Anzahl an Wartepositionen. Sind bei Eintreffen eines Anrufs alle Wartepositionen belegt, so erhält der Anrufer ein Besetztzeichen.

Die ACD-Anlage ist ein zentraler Bestandteil jedes IB Call Centers. Sollen Anrufe zwischen verschiedenen Standorten verteilt werden, wird eine Verbindung zwischen den Teilnetzen der Standorte benötigt. Diese kann beispielsweise über eine VPN-Lösung realisiert werden. Unter VPN wird hierbei ein virtuelles Netz verstanden, das Informationen, unter Verwendung eines speziellen Protokolls, zwischen verschiedenen Knotenpunkten austauscht. Eine weitere ACD-Anlage oder ein VPN-Router fungieren

in einem derartigen System als zentraler Verteilerknoten, der die Anrufe nach einer festgelegten Strategie an die ACD-Anlagen der einzelnen Standorte weiterleitet.

3.1.4 Outbound

Im Gegensatz zum Inbound gehen die Gespräche im Outbound (OB) vom Call Center selbst aus. Die Agents arbeiten dabei Adresslisten ab, die jeweils auf bestimmte Kundengruppen oder Themen zugeschnitten sind. Die Abfolge der Adressabarbeitung und das Anwählen der Rufnummern übernimmt dabei ein sogenannter Dialer. Der Dialer ist eine Software die automatisch anwählt und den Anruf, bei erfolgreicher Verbindung, an einen freien Agent mit entsprechendem Skill weiterleitet. OB wird vor allem für Marktforschung und Umfragen eingesetzt, aber auch zunehmend für den direkten Vertrieb. Ein Beispiel für eine mögliche Aufgabe eines OB Call Centers ist die Vermittlung von Handy- oder Festnetzverträgen. Da Best Service Routing nur für den IB Verwendung findet, soll in dieser Arbeit nicht weiter gesondert auf den OB eingegangen werden.

3.2 Wachstum als Leitmotiv der Call Center Branche

Die deutsche Wirtschaft brummt, heißt es seit einiger Zeit wieder. Und obwohl die großen Vorzeigekonzerne wie VW, Siemens, BMW oder die Deutsche Telekom AG meist nur mit Negativschlagzeilen über Massenentlassungen von sich reden machen, sinken die Arbeitslosenzahlen insgesamt doch. Ein wesentlicher Motor des von Medien und den Politikern verkündeten Aufschwungs ist dabei vor allem in der Dienstleistungsbranche zu suchen. Die größten Zuwächse finden sich besonders bei unternehmensnahen Dienstleistern wie Werbeagenturen, Unternehmensberatungen und natürlich Call Centern.

Die Call Center Branche in Deutschland verzeichnet über die letzten Jahre ein stetiges Wachstum bei Umsatz und Anzahl der Beschäftigungsverhältnisse. Trotz neuerer gesetzlicher Regelungen, die sich beispielsweise gegen unaufgeforderte Telefonwerbung richten, existiert weiterhin ein positiver Trend, der zeigt, dass auch in den nächsten Jahren mit einer steigenden Nachfrage nach telefonischen Dienstleistungen zu rechnen ist. Der daraus resultierende Bedarf nach fähigen aber auch kostengünstigen Arbeitskräften kann vielerorts kaum noch gedeckt werden, weshalb neue Call Center heute vorzugsweise in strukturschwachen Regionen und zunehmend auch im osteuropäischen Ausland angesiedelt werden. Da die angebotenen Stellen

häufig in den Bereich des so genannten Niedriglohnsektors fallen, ist eine regional hohe Verfügbarkeit von Arbeitsuchenden meist ein Vorteil und auch die steigende Zahl privater Arbeitsvermittlungen und Zeitarbeitsfirmen leisten einen nicht zu verachtenden Beitrag zur Sicherung des Nachschubs an der Ressource Mensch.

Viele Unternehmen wie zum Beispiel die Deutsche Telekom AG geben heutzutage ihr Tele-Dienstleistungsgeschäft an spezialisierte Unternehmen ab, mit denen sie vertragliche Vereinbarungen, hinsichtlich der zu erbringenden Leistung und der monetären Gegenleistung, treffen. Gründe für dieses Business Process Outsourcing (BPO), also die Auslagerung von Geschäftsprozessen, sind meist Überlegungen zu Kosteneinsparungen. Die auf Tele-Dienste spezialisierten Call Center Unternehmen können dieselben Dienstleistungen in den meisten Fällen wesentlich kostengünstiger erzeugen als Unternehmen bei denen dies nicht zum eigentlichen Kerngeschäft gehört.

Ein wichtiger Gesichtspunkt, der beispielsweise bei der Konzeption eines neuen Call Centers berücksichtigt werden muss, ist die Tatsache, dass Call Center als Economies of Scale angesehen werden können. Das bedeutet, dass sie mit zunehmender Größe wirtschaftlicher arbeiten, da beispielsweise die Auslastung der Agenten erhöht wird. Ein sehr anschauliches Beispiel hierfür ist die Vorstellung, dass ein mittleres Call Center mit 50 Agents in 5 Einheiten aufgeteilt wird. Jedes der nun entstandenen kleinen Call Center erhält 10 Agents und genau ein Fünftel der Anrufe zugeleitet. Es kann nun passieren, dass ein Anruf in einem Call Center ankommt, dessen Queue schon voll belegt ist oder in der sich zumindest noch wartende Anrufe befinden, obwohl in einem der anderen Call Center noch unausgelastete Agents warten. Der Anrufer würde durch das Besetztzeichen abgewiesen werden oder müsste warten, obwohl eigentlich noch freie Kapazitäten in den vier verbleibenden Call Centern verfügbar wären.

Mit einer zentralen Warteschlange würde das nicht passieren, da ein ankommender Anruf, sofern irgendwo freie Agenten sitzen, auch an diese geleitet wird. Durch Wiedervereinigung der fünf Call Center verringert sich also die Wahrscheinlichkeit, dass Kunden abgewiesen werden, weiterhin wird die durchschnittliche Wartezeit verkürzt und die Auslastung der Agents erhöht.

Mit Hilfe von Best Service Routing ist es nun möglich, mehrere Call Center so zu vernetzen, dass ein virtuelles Super-Call-Center entsteht. Jeder verbundene Standort besitzt dabei zwar eine eigene Warteschlange, aber das System ist darauf ausgerichtet, alle Warteschlangen möglichst gleich stark zu belasten. Auf diese Weise wird angestrebt, dass sie sich in ihrer Gesamtheit so ähnlich verhalten wie eine einzige Warteschlange mit der Summe der Wartepositionen der Einzelqueues. Anrufer erhalten erst dann das Besetztzeichen, wenn sämtliche Warteplätze aller Standorte besetzt sind.

Da Größenvorteile bei entsprechender Auftragslage auch Wettbewerbsvorteile bedeuten, ist zu erwarten, dass standortübergreifende Routingstrategien wie Best Service Routing auch zukünftig weiter an Bedeutung gewinnen werden.

3.3 Die Personaleinsatzplanung

3.3.1 Einflußgrößen der Einsatzplanung

Ein Schlüssel, der im Call Center Betrieb sehr stark über den wirtschaftlichen Erfolg oder Misserfolg entscheidet, ist eine effektive Personaleinsatzplanung. Dies folgt aus dem überdurchschnittlich hohen Personalkostenanteil, der ca. 70% der Gesamtkosten betragen kann.

Während die OB-Planung hauptsächlich durch im Tagesverlauf schwankenden Erreichbarkeiten bei privaten Telefonanschlüsse sowie durch Start- und Endtermine der OB-Kampagnen bestimmt ist, steht man im klassischen IB-Betrieb zu jedem Zeitpunkt einem konkreten Anrufaufkommen gegenüber. Dieses muss grob gesagt von einer passenden Anzahl Agents bewältigt werden. Das Entscheidende ist hierbei, für jeden Zeitpunkt auch die wirtschaftlich sinnvollste Anzahl an Agents einzuplanen. Werden zu viele Agents eingesetzt, so entstehen Bereitzeiten und damit unnötige Kosten, denn die Agents müssen bezahlt werden, auch wenn sie unproduktiv warten. Sind zu wenige Agents eingesetzt, so wirkt sich dies in einem schlechten Service aus, der dadurch entsteht, dass Kunden im Durchschnitt wesentlich länger warten müssen, bis sie mit einem freien Agent verbunden werden.

Im IB Vertriebs-Call-Center spielen Wartezeiten auf Kundenseite meist eine untergeordnete Rolle. Im Normalfall werden dort durch Maximierung der durchschnittlichen Deckungsbeiträge pro Mitarbeiterstunde auch die Gewinne des Call Centers maximiert. Während also in Dienstleistungs- und Betreuungs-Call-Centern die angestrebte Servicequalität, die mittlere Bedienzeit und die Ankunftsrate der Anrufe die entscheidenden Basiswerte für eine kostenminimale Bedarfsermittlung zur Erreichung von vertraglichen Vorgaben sind, sollte der Mitarbeitereinsatz im IB Vertriebs-Call-Center zu jedem Zeitpunkt direkt auf die Maximierung der Gewinne hin optimiert werden. Die Basiswerte sind dann neben der durchschnittlichen Ankunftsrate der Anrufe auch die Auslastung der Agents und vor allem der durchschnittliche Umsatz pro Anruf. Hohe Wartezeiten können sich natürlich auch im Vertriebs-Call-Center negativ auf die Gewinne auswirken, da dadurch die Wahrscheinlichkeit erhöht wird, dass immer mehr Kunden schon vorher auflegen und somit nichts mehr zum Umsatz beitragen.

Der Einsatz von Best Service Routing ist wie schon beschrieben auf den IB beschränkt und aus der Erfahrung heraus dort hauptsächlich für Dienstleistungs- und Vertriebs-Call-Center gebräuchlich. Im Weiteren soll deshalb nur zwischen diesen beiden Arten differenziert werden. Betreuungs-Call-Center sind aber, sofern sie eine Servicequalität vereinbart haben, von ihren Planungsanforderungen her mit Dienstleistungs-Call-Centern gleichzusetzen.

3.3.2 Der Servicelevel

Der Servicelevel ist ein Index, der die Servicequalität eines IB Call Centers misst. Mit ihm kann überprüft werden, ob das Call Center eine vertraglich festgelegte Mindestgrenze der Servicequalität nicht unterschreitet.

Klassischerweise ist der Servicelevel eine fortwährend gemessene Größe, die angibt, wie viel Prozent der Kunden mit einer Wartezeit unterhalb eines festgesetzten Zeitlimits bedient wurden. Kunden, die vor Erreichen der Zeitgrenze auflegen, werden dabei im Normalfall nicht mit eingerechnet.

Im Allgemeinen bieten Call Center ihren Service im Auftrag von Dritten an, mit denen das Call Center Unternehmen entsprechende Verträge hat. Zumindest für Dienstleistungs- und Betreuungs-Call-Center wird hierbei ein sogenanntes Servicelevel-Agreement festgelegt, das beispielsweise beinhalten kann, dass 80 Prozent der ankommenden Anrufe nicht länger als 20 Sekunden warten dürfen, bis sie von einem Agent bedient werden. Dieses Beispiel ist gleichzeitig einer der am häufigsten in der Praxis anzutreffenden Servicelevel der allgemein als Servicelevel 80/20 bezeichnet wird. Hält ein Call Center die Servicelevelnorm nicht ein, so drohen ihm entweder Vertragsstrafen oder Prämienabzüge. Es ist also im Interesse des Call Center Betreibers den Mindestwert nicht zu unterschreiten.

Bei Existenz eines Servicelevel-Agreements müssten die Agents optimalerweise so geplant werden, dass die Servicelevelgrenze zu jedem Zeitpunkt des Operativbetriebs genau getroffen wird. Da in der Praxis allerdings meistens einige Mitarbeiter durch beispielsweise Krankheit ausfallen, sollte immer mit einem gewissen Sicherheitspuffer geplant werden.

3.3.3 Workforce-Management-Systeme

Softwareanwendungen für die Planung der benötigten Besetzungsstärke sowie der zeitmäßigen Einteilung des Personals werden als Workforce-Management-Systeme bezeichnet. Sie bestehen üblicherweise aus verschiedenen Teilanwendungen oder Modulen. Neben der eigentlichen Plananwendung, mit der letztendlich die Arbeitsplatzpläne erzeugt werden, benötigt ein Call Center, zumindest wenn es Inbound betreibt, einen sogenannten Forecast. Dieser Forecast ist das Prognoseinstrument für das Anrufaufkommen und im Allgemeinen auch für die Berechnung des daraus resultierenden Mitarbeiterbedarfs zuständig. Historische Daten, die auf einer ACD-Anlage gemessen wurden, sind hierbei die Grundlage der Prognose.

Die vom Forecast erzeugten Bedarfskurven dienen der Planungsanwendung als wichtigste Orientierung bei der Vergabe der Einsatzzeiten an die Mitarbeiter. Neben den Bedarfskurven benötigt der Plan natürlich noch eine große Menge anderer Daten wie Schichtmodelle, spezifische Mitarbeiterinformationen oder allgemein geltende Regeln, wie beispielsweise arbeitsrechtliche Bestimmungen.

Abbildung 2.1 gibt eine kleine Übersicht über die wichtigsten Komponenten und grundlegende Daten, die in den meisten Workforce-Management-Systemen anzutreffen sind. Die hier aufgezeigten Komponenten müssen nicht zwangsläufig eigenständige Anwendungen sein. In der Praxis vertreiben kommerzielle Anbieter ihre Produkte aber meist in Form einzelner Module, die bei Bedarf extra hinzugekauft werden müssen.

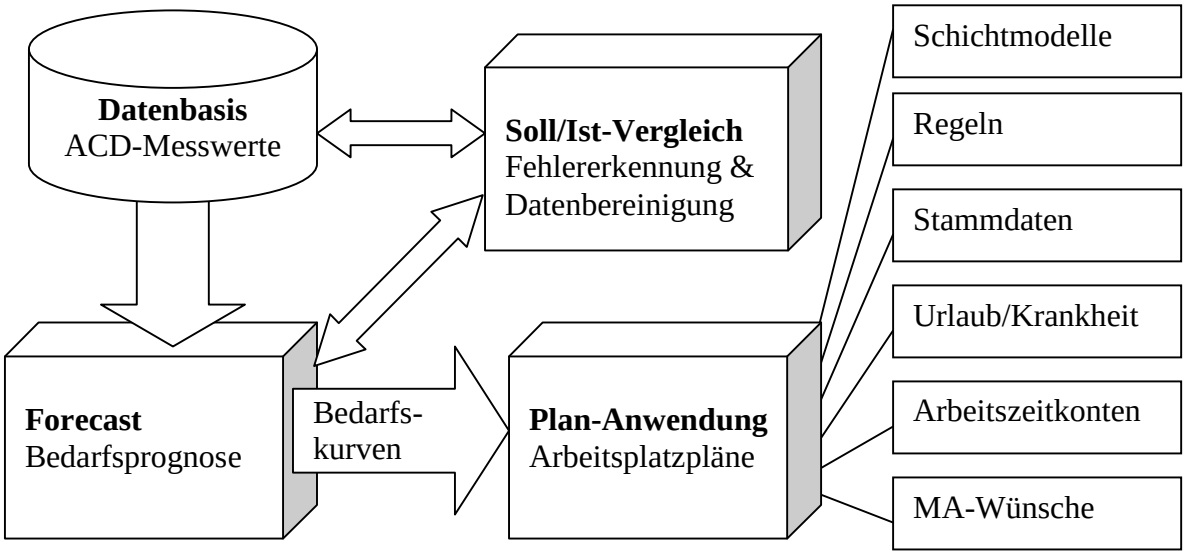


Abb. 3.1: Komponenten von Workforce-Management-Systemen

Call Center Agents arbeiten üblicherweise in Schichten. Ein angewendetes Schichtmodell sollte aber die Arbeitszeit nicht starr festlegen, sondern einen Zeitrahmen liefern in dem sie mit einem gewissen Spielraum positioniert werden kann. Da es während der Planung darauf ankommt, die vom Forecast erzeugten Bedarfskurven möglichst genau abzudecken, verbessert eine höhere Variabilität beim Arbeitseinsatz auch das Planungsergebnis.

Die für die Planungsqualität bestimmende Komponente ist der in der Plananwendung implementierte Planungsalgorithmus. Da die Planerstellung ein Optimierungsproblem ist, für das kein allgemeingültiges Lösungsverfahren existiert, werden unter Beachtung der Schichtplänen und weiterer jeweils geltender Regeln bestimmte Heuristiken und meistens auch evolutionäre Algorithmen eingesetzt, um sich dem bestmöglichen Plan so gut es geht anzunähern.

Mit einem evolutionären Algorithmus ist in diesem Zusammenhang ein Softwareprogramm gemeint, dass versucht, die Gesetze der natürlichen Auslese nachzuahmen. Dabei werden Populationen unterschiedlicher Pläne erzeugt und nach vorgegebenen Regeln bewertet. Die dabei am besten bewerteten Pläne gehen wiederum in eine neue Population über (Selektion). Die neue Population erhält weiterhin neu kombinierte Pläne (Rekombination), sowie zufällig gestreute Änderungen (Mutationen). Durch mehrfache Wiederholung des Vorgangs kann, bei richtiger Gewichtung der Bewertungskriterien und der Parameter für Selektion, Rekombination und Mutation, eine sehr hohe Planungsqualität erzielt werden. Vielfach ist aber gerade das Finden der besten Einstellungen so schwierig, das dafür im Prinzip ebenfalls ein optimierender Algorithmus notwendig wäre.

Workforce-Management-Systeme benötigen neben dem Forecast und der Plananwendung eine zusätzliche Analysekomponente, um einen fortlaufenden Abgleich zwischen den Prognosewerten und den real eintretenden Ereignissen durchführen zu können. Der Soll-Ist-Vergleich überprüft die ACD-Werte auf eventuell auftretende überproportionale Abweichungen des realen Anrufaufkommens und des Mitarbeiterbedarfs von der Vorhersage. Die Abweichungen werden bewertet und je nach Ausprägung werden entsprechende Maßnahmen eingeleitet. Hauptsächlich dient der Soll-Ist-Vergleich zur Qualifizierung der Messwerte für zukünftige Prognosen. Wurde eine Unregelmäßigkeit identifiziert, so wird das entsprechende Intervall der Messwerte für die Verwendung in kommenden Prognosen entweder niedriger gewichtet oder als komplett nicht verwertbar markiert.

3.3.4 Prognoseanforderungen beim Best Service Routing

Für ein Prognosesystem des Best Service Routings wird neben dem Forecast und Soll-Ist-Vergleich eine zusätzliche Komponente benötigt werden, die neben der Qualifikation auch eine Transformation einiger Messwerte durchführen muss. Durch die Transformation sollen Wertreihen generiert werden, welche in etwa die theoretisch optimale Besetzung und die damit in Beziehung stehende theoretisch erzielbare Anrufmenge am eigenen Standort für den gemessenen Zeitraum wiedergeben. Diese Wertreihen sind dann die wichtigste Grundlage der neuen Prognose des Mitarbeiterbedarfs, allerdings nicht die einzige. Über die Abweichung des theoretischen Optimalwertes des Mitarbeitereinsatzes vom tatsächlich am eigenen Standort registrierten, können indirekt auch Aussagen über die Dynamik des Gesamtsystems abgeleitet werden, da eine Unterbesetzung am eigenen Standort wegen der ausbalancierenden Wirkung des Best Service Routings auch eine Unterbesetzung an den anderen Standorten bedeutet. So kann in die Prognose des zukünftigen Mitarbeitereinsatzes beispielsweise einfließen, welche Reaktionen der anderen Standorte auf Abweichungen vom theoretischen Optimum zu erwarten sind. Es könnte beispielsweise prognostiziert werden, dass nach Tagen mit starker Unterbesetzung in Folgewochen mit einer überproportional hohen Besetzungsverstärkung an den anderen Standorten gerechnet werden muss.

Das Hauptproblem wird aber darin bestehen, die Zielfunktion für die nachträgliche Bestimmung des optimalen Mitarbeitereinsatzes zu finden. Die an den einzelnen Standorten gemessenen Anrufaufkommen und Wartezeiten lassen ohne Wissen über die Konfiguration der übrigen Call Center keine verlässliche Aussage darüber machen, wie viele Anrufe für eine optimale Auslastung möglich gewesen wären. Die einzigen Indizien für diese Konfiguration sind die Menge der insgesamt ankommenden Anrufe und der Anteil, der davon den eigenen Standort erreicht hat. Durch Best Service Routing werden dem einzelnen Call Center zu jedem Zeitpunkt maximal ja nur so viele Anrufe zugeleitet, wie es mit seiner aktuellen Besetzungsstärke verkraften kann. Bei zu hohem Anrufaufkommen werden die überzähligen Anrufe am zentralen Verteiler abgeblockt. Das Abblocken ist aber eher ein Extremfall, der nur bei wirklich hoher Unterbesetzung eintritt. Im Normalfall wirkt sich eine Unterbesetzung so aus, dass die Wartezeiten der Kunden nach oben gehen und der Anteil der Aufleger steigt. Die Anzahl der auftretenden Besetztfälle und die Höhe der jeweils gemessenen durchschnittlichen Wartezeiten der Anrufer, lassen Rückschlüsse auf den Grad der Unterbesetzung im Gesamtsystem zu. Wie hoch die Unterbesetzung allerdings für den einzelnen Standort anzusetzen ist, hängt wie erwähnt von der Besetzungskonfiguration des Gesamtsystems ab. Dasselbe gilt auch für Intervallen, in denen hohe Bereitzeiten

und demnach ein Mitarbeiterüberschuss gemessen wurden. Aus den Messwerten ist auch hier nicht eindeutig ersichtlich, wie hoch der Überschuss am eigenen Standort tatsächlich ist.

Die zusätzliche Komponente muss also, durch Anwendung eines Optimierungsmodells sowie unterschiedlicher Analysemethoden, eine Art Korrektur des am Standort gemessenen Anrufaufkommens und Mitarbeitereinsatzes vornehmen. Die Korrekturwerte müssen in etwa wiedergeben, wie die optimale Mitarbeiterkonfiguration ausgesehen hätte, um die Gewinne zu maximieren bzw. die Kosten zu minimieren. Die Erstellung der Korrekturwerte basiert dabei hauptsächlich auf der Auswertung von Messwerten wie der durchschnittlichen Wartezeit der Anrufer, den Besetztfällen, der ankommenden Anrufe im Gesamtsystem und am eigenen Standort sowie den eventuell gemessenen Bereitzeiten der Agents.

Die Aufgabe ist es also, ein mathematisches Grundmodell zu entwickeln, das allgemeine Verteilungsregeln des Best Service Routings berücksichtigt und das auf Grundlage der genannten Messwerte eine nachträgliche Optimierung vornimmt und so die Voraussetzung für eine Prognose liefert, die auch die Reaktionen der Mitbewerber einkalkuliert. Die Verteilungsregeln selbst und das grundlegende Verhalten eines Best Service Routing Systems müssen dafür natürlich im Vorfeld dieser Entwicklung analysiert werden, um dadurch die geltenden Gesetzmäßigkeiten bestimmen zu können.

Der Soll-Ist-Vergleich könnte in diesem System, neben der Erledigung seiner im letzten Abschnitt beschriebenen Standardaufgaben zusätzlich noch der Regulierung und Optimierung des beschriebenen Korrekturprozesses dienen. Beispielsweise ist ein Lernalgorithmus denkbar, der die jeweils erreichte Prognosequalität anhand der operativ erreichten Ergebnisse bewertet, und darauf aufbauend Änderungen an bestimmten Parametern der Zielfunktion vornimmt. So könnte diese in Richtung einer besseren Lösung weiterentwickelt werden.

4 Die Warteschlangentheorie

4.1 Begriffsabgrenzung

Vor einer tieferen Analyse der Vorgänge die speziell in einem Call Centern mit Best Service Routing ablaufen, soll an dieser Stelle darauf aufmerksam gemacht werden, dass IB Call Center allgemein von stochastischen Prozessen geprägt sind. Sowohl die Abstände in denen Anrufe hereinkommen als auch die Dauer der Anrufe, sind im Einzelnen nicht vorhersehbar. Sie unterliegen zufälligen Abweichungen von einem konkreten Erwartungswert, der aber im Mittel relativ konstant bleibt.

In der Modellvorstellung baut ein IB Call Center auf einem Warteschlangenmodell auf. Die zugrunde liegende Warteschlangentheorie gehört zur angewandten Wahrscheinlichkeitstheorie und beschäftigt sich mit der mathematischen Betrachtung von Bediensystemen, bei denen Aufträge an einzelnen Bedienstationen bearbeitet werden. Enthält das System daneben noch einen Warteraum, so wird von einem Wartesystem gesprochen, andernfalls von einem Verlustsystem.

Im IB Call Center repräsentieren die von Agents besetzten Arbeitsplätze die Bedienstationen und die dort bearbeiteten Aufträge sind telefonische Dienstleistungen. Die Anzahl der Wartepositionen in der Queue eines Call Centers ergeben sich aus der Anzahl vorhandener Telefonleitungen minus der Anzahl besetzter Arbeitsplätze. Die Queue dient dazu, die ungleich verteilten Zwischenankunftszeiten und Bediendauern der Anrufe auszugleichen und hat demnach eine Pufferfunktion.

Abbildung 3.1 zeigt ein Wartesystem, das als Modell eines vereinfachten Call Centers angesehen werden kann. Zur Abbildung realer Call Center ist dieses Modell jedoch relativ ungeeignet, da es weder Aufleger noch Besetztfälle berücksichtigt.

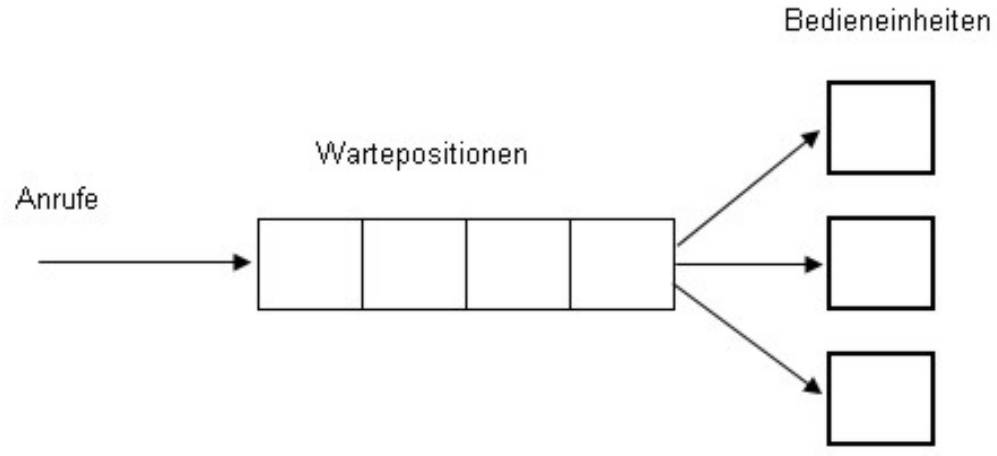


Abb. 4.1: Ein einfaches Wartesystem

Sind mehrere Wartesysteme miteinander verbunden, so spricht man von einem Warteschlangennetz. Die einzelnen Wartesysteme können dabei parallel oder in Reihe geschaltet sein. Allgemein wird zwischen geschlossenen und offenen Warteschlangennetzen unterschieden. Geschlossene Netze haben keine externen Zu- oder Abgänge. Da in Call Centern Anrufe im System eingehen und dieses auch wieder verlassen, handelt es sich demzufolge um offene Warteschlangennetze. Ein System mit mehreren in Beziehung stehenden Call Centern, wie beim Best Service Routing, erhöht natürlich die Komplexität des Ausgangsmodells aus Abbildung 4.1 erheblich, da viel mehr Parameter, wie unterschiedliche durchschnittliche Bedienzeiten und unterschiedlich große Warte- und Bedienräume zu berücksichtigen sind. Außerdem finden sich in der Literatur bisher keine oder nur wenige verlässliche Informationen geschweige denn Formelwerke zu den dort auftretenden Effekten und Wechselbeziehungen.

4.2 Littles Gesetz

Littles Gesetz ist eine der wichtigsten Gesetzmäßigkeiten der Warteschlangentheorie. Es besagt, dass die durchschnittliche Anzahl von Aufträgen in einem Warteschlangenmodell gleich dem Produkt ihrer mittleren Ankunftsrate und ihrer mittleren Durchlaufzeit ist.

$$L = \lambda * W$$

Das Gesetz wurde 1961 von John D.C. Little bewiesen. Es bezieht sich immer auf ein eingeschwungenes System, in dem die Menge der eintreffenden Aufträge auch bewältigt werden kann. Es ist unabhängig von der jeweils vorherrschenden Verteilungsform der Zwischenankunftszeiten und Durchlaufzeiten und deshalb allgemeingültig. Littles Gesetz ist ein wichtiger Baustein bei der Lösung vieler mathematischer Probleme der Warteschlangentheorie.

4.3 Die Kendall-Notation

Unterschiedliche Ausprägungen von Wartesystemen lassen sich mit der Kendall-Notation beschreiben. Diese geht auf den Mathematiker David George Kendall zurück, und fordert neben Angaben über die Kapazität des Systems und Anzahl der Wartepositionen auch Angaben zur statistischen Verteilung der zufallsbehafteten Größen wie etwa der Anrufdauer.

Die Notation besteht aus einer Zeichenkette der Form (.)/(.)/c/K. Die erste Position steht hierbei für die statistische Verteilung des Ankunftsprozesses, also der Zwischenankunftszeiten. An zweiter Stelle wird die Verteilung der Bedienzeiten festgehalten. Weiterhin steht c für die Anzahl gleicher Serviceeinheiten, also die Arbeitsplätze, und K für die Anzahl der Telefonleitungen.

Mögliche Symbole für Verteilungsformen sind M, D und G, wobei M für eine Exponentialverteilung, D für eine rein deterministische Verteilung und G für eine allgemeine oder beliebige Verteilung steht. Vielfach finden sich in der Literatur noch weitere Symbole für andere Verteilungsformen sowie Erweiterungen der Notation um zusätzliche Positionen. So kann beispielsweise die Ungeduld wartender Anrufer mit einbezogen werden. Üblicherweise wird das zusätzliche Symbol für die Verteilung des Auflegeverhaltens mit einem Pluszeichen an die beschriebene Zeichenkette angefügt. (vgl. Helbert (2003), S. 34 - 37)

Ein Modell, das die realen Gegebenheiten eines Call Centers schon relativ gut beschreibt, ist beispielsweise folgendes: (M/M/c/K + M).

Sowohl die jeweilige Anrufdauer als auch die Zwischenankunftszeiten und die Ungeduld der Anrufer sind im Einzelnen nicht vorhersagbar und unterliegen zufälligen Schwankungen. Alle drei Merkmale haben aber, in Abhängigkeit von der Tageszeit und der Menge ankommender Anrufe, jeweils eine bestimmte Wahrscheinlichkeitsverteilung mit einem konkreten Erwartungswert. In der Modellannahme geht die eben angeführte Notation in allen Fällen von einer Exponentialverteilung aus, die durch die drei M symbolisiert wird. Exponentialverteilungen stimmen zwar nicht unbedingt mit den realen Verteilungen überein, kommen ihnen aber relativ nahe und sind vor allem im praktischen Einsatz noch recht gut zu handhaben, da mit ihnen vergleichsweise einfach gerechnet werden kann.

4.4 Markov-Prozesse

Markov-Prozesse sind eine grundlegende Klasse stochastischer Modelle. Während die Bezeichnung Markov-Prozess hauptsächlich für den stetigen Fall, also einen kontinuierlichen Zeitverlauf, verwendet wird, spricht man im diskreten Fall (Folge von Zeitpunkten) von Markov-Ketten. Im Folgenden soll allgemein die englische Bezeichnung Markov-Chain (MC) verwendet werden.

Im Zusammenhang mit Warteschlangenmodellen soll weiterhin nur von MCs erster Ordnung im zeitdiskreten Fall mit endlichen Zustandsräumen ausgegangen werden. MCs erster Ordnung beschreiben Systeme, bei denen die Zukunft nur vom aktuellen Systemzustand bestimmt wird und keine Kenntnis der vorangegangenen Zustandsfolge nötig ist. Neben den möglichen Zuständen beschreibt ein MC ein System über die Übergangswahrscheinlichkeiten zwischen diesen Zuständen und eine Anfangsverteilung. Ein MC besteht also aus einer endlichen Zustandsmenge E = {1,2,…,ℓ}, mit allen möglichen Systemzuständen, einem Vektor α, der für jeden Zustand in E dessen Wahrscheinlichkeit im Zeitpunkt Null enthält und einer Matrix P, mit allen Übergangswahrscheinlichkeiten.

Wenn E also ℓ Elemente beinhaltet, enthält α auch ℓ Wahrscheinlichkeiten. Für jedes $i \in E$ ist α_i die Wahrscheinlichkeit, dass sich das System zum Zeitpunkt t=0 im Zustand i befindet und es gilt: $\alpha_i \in [0;1], \sum_{i=1}^{\ell} \alpha_i = 1$

P ist eine ℓ× ℓ Matrix. Für jedes Paar $i, j \in E$ ist $p_{i,j}$ die bedingte Wahrscheinlichkeit, dass das System ausgehend von Zustand *i* in den Zustand *j* übergeht. Es gilt dabei:

$$p_{ij} \in [0;1], \sum_{j=1}^{\ell} p_{ij} = 1$$

Eine Folge Zufallszahlen $X_0, X_1, .., X_n$ mit Werten in E heißt Markov-Kette erster Ordnung mit der Startverteilung α_i, wenn eine Matrix existiert, so dass:

$$P(X_n = i_n \mid X_{n-1} = i_{n-1}, ..., X_0 = i_0) = P(X_n = i_n \mid X_{n-1} = i_{n-1}) = p_{i_{n-1} i_n}$$

Dies bezeichnet lediglich die beschriebene Eigenschaft der Matrix, für jeden Zustand die bedingten Wahrscheinlichkeiten für den Übergang zu den anderen Zuständen anzugeben. Die Zukunft hängt dabei immer nur vom aktuellen Zustand ab.

Ein Warteschlangenmodell lässt sich beispielsweise durch einen so genannten Geburts- und Sterbeprozess darstellen. Abbildung 3.2 zeigt diesen für ein einzelnes Call Center. Die Pfeile stellen die möglichen Zustandsübergänge dar, die durch neu eintreffende oder abgehende Anrufe ausgelöst werden. Im Zustand 0 befinden sich weder Anrufe in der Warteschlange noch in Bedienung. Im Zustand K sind alle Bedienplätze ausgelastet und auch alle möglichen Warteposition besetzt. Zustand 0 und K besitzen jeweils nur eine Übergangsmöglichkeit während alle anderen Zustände immer jeweils zwei besitzen.

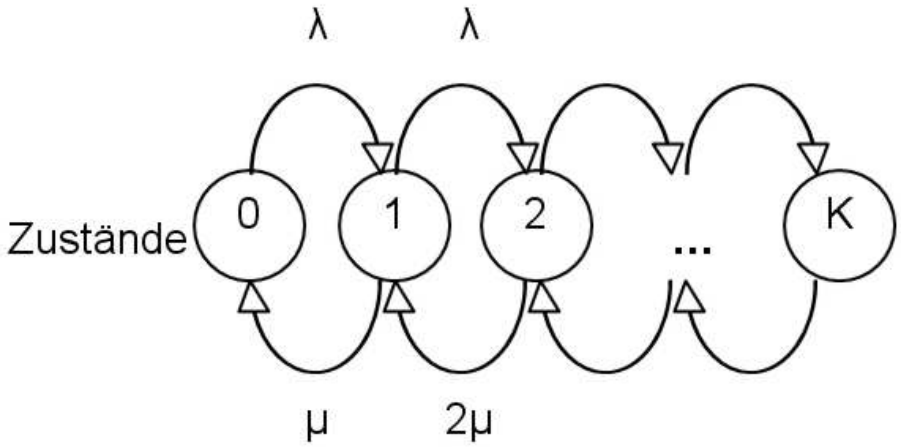


Abb. 4.2: Geburts- und Sterbeprozess

Ein Schritt nach rechts bedeutet im Bild jeweils, dass ein neuer Anruf in das System gelangt, während ein Schritt nach links durch das Abgehen eines Anrufs, also durch erfolgreiche Bedienung oder Auflegen, ausgelöst wird.

4.5 Die praktische Anwendung der Warteschlangentheorie

Die Warteschlangentheorie findet überall dort Anwendung, wo gleichartige Aufträge von Bedienstationen abgearbeitet werden müssen. Ein Beispiel hierfür ist die Bearbeitung von Tasks durch die CPU eines Computers. Eine CPU kann immer nur eine Rechenoperation gleichzeitig ausführen, während alle anderen Aufträge warten müssen. Auch wenn hierbei durch Scheduling-Algorithmen der Eindruck von Multitasking erzeugt wird, kann wirkliche Parallelität nur mit mehreren CPUs erreicht werden. Rechnerarchitekturen sind ein wichtiges Anwendungsgebiet der Warteschlangentheorie, aber die zunehmende wirtschaftlichen Bedeutung der Tele-Dienstleistungsbranche, macht auch Call Center zu einem wichtigen Gegenstand der Forschung auf diesem Gebiet. Mit Hilfe von Optimierungsverfahren aus dem Operations Research (OR) kann, ausgehend von Warteschlangenmodellen, beispielsweise auf die wirtschaftlichste Kapazitäten eines Call Centers geschlossen werden. Die wichtigsten Ausgangsgrößen sind dabei die angestrebte mittlere Bedienzeit, das erwartete Anrufaufkommen und die angestrebte Servicequalität. Es existieren unterschiedliche Ansätze, die hauptsächlich aufgrund unterschiedlicher Ausgangsmodelle zustande kommen.

Eine der ersten Anwendungen der Warteschlangentheorie entwickelte der dänische Mathematiker Agner Krarup Erlang schon 1909 im Auftrag einer Telefongesellschaft. Das seiner Arbeit zugrunde liegende Warteschlangenmodell lässt sich durch die Notation (M/M/c/∞) beschreiben. Das bedeutet, dass es von einigen in der Praxis nicht

zutreffenden Gegebenheiten wie etwa einer unbegrenzten Anzahl an Wartepositionen und einer unendlichen Geduld der Anrufer ausgeht. Nichtsdestotrotz werden einige seiner Berechnungsvorschriften, wie etwa die Erlang-C-Formel, auch heute noch in kommerziellen Anwendungen eingesetzt, um beispielsweise die Anzahl benötigter Agents bei einem vorgegebenen Servicelevel abzuschätzen.

Formeln, die auf weniger abstrahierenden Modellen basieren, liefern mittlerweile aber schon weitaus bessere Ergebnisse. Allerdings müssen, aufgrund der Komplexität der Formeln, für die Lösungsfindung meist lineare oder sogar dynamische Optimierungsverfahren angewendet werden. Zunehmend werden auch Simulationssysteme eingesetzt um sich einer optimalen Einzellösung noch weiter anzunähern oder bestimmte Parameter eines Warteschlangenmodells abzuschätzen. Auch lernfähige Algorithmen wie künstliche neuronale Netze können zu diesem Zweck eingesetzt werden. Die Wahl der jeweiligen Methoden hängt hauptsächlich vom Zeitaufwand bzw. vom Maximalwert des tolerierten Prognosefehlers ab.

5 Zeitreihenanalyse

5.1 Zeitreihen und stochastischen Prozessen

Der Begriff der Zeitreihe ist eng mit dem Begriff des stochastischen Prozesses verknüpft und letztendlich werden beide in der Literatur häufig synonym verwendet. Der Hauptunterschied ist lediglich in der Zielausrichtung des jeweils zugrunde liegenden mathematischen Gebietes zu finden. Während die Zeitreihenanalyse in das Gebiet der Statistik fällt und versucht, Modelle zur Prognose zeitlich geordneter Daten, den sogenannten Zeitreihen, aufzustellen, stehen in der Stochastik die Strukturen von Zufallsfunktionen, durch welche sich stochastische Prozesse beschreiben lassen, im Vordergrund. Ein wichtiger Begriff im Zusammenhang mit stochastischen Prozessen spielt aber auch für die Zeitreihenanalyse eine wichtige Rolle. Die Rede ist von der sogenannten Stationarität stochastischer Prozesse. Stationarität bedeutet, dass die Zeitreihe oder die Zufallsfunktion zu jedem Zeitpunkt denselben Erwartungswert und dieselbe Varianz aufweisen. Nichtstationarität bedeutet bei der Zeitreihenanalyse, dass ein Trend für das betrachtete Merkmal oder für die Varianz des betrachteten Merkmals vorliegt.

5.2 Überblick zu Ansätzen der Prognosebildung

Um aus den Zeitreihen verwertbare Prognosen für beispielsweise das Anrufaufkommen zu gewinnen, werden üblicherweise statistische Methoden der Zeitreihenanalyse angewendet. Daneben verwenden einige Softwaresysteme Methoden des maschinellen Lernens bzw. der künstliche Intelligenz, um etwa charakteristische Verteilungsmuster oder Ausreißer in den Zeitreihen zu identifizieren, aber teilweise auch um komplette Vorhersagen zu machen. Verschiedene Implementierungen künstlicher neuronaler Netze (KNN) stellen beispielsweise ein flexibles, wenn auch im Hinblick auf den Prognosefehler nicht sehr genaues Mittel dar, um Prognosen auf Basis von kurzfristigen historischen Werten zu erzeugen. Es existieren verschiedene Varianten von KNN die auf unterschiedlichen Abstraktionsstufen den biologischen Strukturen der Nervenzellen eines Gehirns nachempfunden sind. Mit KNNs können beliebige Funktionen mit beliebig vielen Eingangs- und Ausgangsgrößen erlernt werden. So könnte beispielsweise ein spezielles KNN mit fünf Eingangs-Neuronen und einem Ausgangs-Neuron lernen, aus den letzten fünf Zeitreihenwerten den erwarteten Zukunftswert abzuschätzen. Es müsste dafür einfach mit Mustern aus historischen Daten trainiert werden. Das Netz würde dabei die Zuordnung eines Wertes zu seinen fünf Vorgängerwerten erlernen und wäre darüber in der Lage, auch für beliebige neue Eingangswerte eine Abschätzung des Folgewertes zu liefern.

Die Schwierigkeiten dieser Methode liegen letztlich, neben der Wahl einer geeigneten Netzarchitektur, in der Auswahl repräsentativer Trainingswerte. Um aktuelle Prognosen zu erstellen, sollten auch die Trainingwerte möglichst nicht veraltet sein, weshalb das Training kontinuierlich weitergeführt werden muss. Ein zu intensives Training kann aber auch zu einem Effekt führen, der als Übertrainieren bezeichnet wird. Da den realen Daten immer auch ein systemspezifisches Rauschen anhängt, werden meist Verfahren eingesetzt, mit denen Trainingswerte aus Messwerten generiert werden. Die Messwerte werden dafür beispielsweise durch bestimmte Auswahl und Glättungsverfahren aufbereitet oder gleich in charakteristische Verteilungen überführt.

Wesentlich unkomplizierter, in den meisten Fällen aber trotzdem ausreichend genau, sind schon einfache statistische Verfahren, wie etwa gleitende Mittelwerte oder die exponentielle Glättung. Mit etwas aufwändigeren Verfahren, die zum Beispiel Trends oder saisonale Schwankungen berücksichtigen, kann dabei die Genauigkeit je nach Anforderungen weiter gesteigert werden. Zu erwähnen sind in diesem Zusammenhang beispielsweise die exponentielle Glättung zweiter Ordnung, die einen linearen Trend mit einbezieht, sowie die schon wesentlich komplexeren ARIMA-Modelle. die neben Trends auch um die Berücksichtigung von Ausreißern, saisonalen Schwankungen und Feiertagen erweitert werden können.

Im Falle des Best Service Routings wird es aufgrund der erhöhten Unsicherheit notwendig sein, zum Teil auf gemischte Strategien zurückzugreifen. Ausgehend von einem Optimierungsmodell, dass möglichst alle zur Verfügung stehenden Messwerte in die richtige Beziehung setzt, sollten bei der letztendlichen Prognose auch Aussagen zur Messwertstreuung und Prognosepräzision beispielsweise in Form von Konfidenzintervallen bereitgestellt werden. Der Anspruch ist die Schaffung eines flexiblen und zuverlässigen Prognoseinstruments.

5.3 Die exponentielle Glättung zweiter Ordnung

Die exponentielle Glättung zweiter Ordnung ist ein Verfahren, mit dem kurzfristige Prognosen unter der Berücksichtigung eines linearen Trends möglich sind. Wegen seiner Leistungsfähigkeit und der noch relativ geringen Komplexität wird das Verfahren auch heute noch sehr häufig eingesetzt. Für das hier zu erarbeitende Prognoseinstrument des Best Service Routings stellt die exponentielle Glättung zweiter Ordnung eine einfach zu implementierende Möglichkeit dar, um beispielsweise die Vorhersage des Anrufaufkommens für das Gesamtsystem vorzunehmen.

Das Verfahren geht, wie schon der Name vermuten lässt, aus der exponentiellen Glättung erster Ordnung hervor. Ein Parameter, der hierbei vorgegeben werden muss, ist der sogenannte Glättungsfaktor α. Dieser kann zwischen Null und Eins gelegt werden und bestimmt die Stärke, mit welcher der letzte gemessene Beobachtungswert das Prognoseergebnis beeinflusst. Je kleiner der Glättungsfaktor, umso geringer ist der Einfluss des letzten Beobachtungswertes.

Es gilt folgendes Gleichungssystem:

$$y^{*}_{t+1} = 2\hat{y}_{t+1} - \hat{\hat{y}}_t$$

$$\hat{y}_{t+1} = \alpha\, y_t + (1-\alpha)\hat{y}_t \qquad \text{erste Glättung}$$

$$\hat{\hat{y}}_{t+1} = \alpha\, \hat{y}_{t+1} + (1-\alpha)\hat{\hat{y}}_t \qquad \text{zweite Glättung}$$

Die zweite Formel ist hierbei dieselbe wie beim Verfahren der exponentiellen Glättung erster Ordnung.

Die Anwendung dieses Verfahrens ist relativ einfach. Es wird immer bei t=1 begonnen, die Wertereihen der ersten und zweiten Glättung zu erzeugen. Für t=1 werden alle fehlenden Werte durch den ersten Beobachtungswert vertreten.

Ein Vorteil des Verfahrens ist, dass bei konstant gehaltenem α für eine neue Prognose nicht die gesamte Zeitreihe zur Verfügung stehen muss, sondern lediglich der letzte Beobachtungswert und die letzten Werte der ersten und zweiten Glättung.

5.4 Konfidenzintervalle

Das Konfidenzintervall ist ein Begriff aus der Statistik, der einen sogenannten Vertrauensbereich beschreibt. Es liefert das Intervall, in dem ein unbekannter Parameter mit einer vorher festgesetzten Wahrscheinlichkeit liegt. Im Unterschied zur Punktschätzung erhält man durch das Konfidenzintervall also ebenfalls eine Information zur Präzision und Zuverlässigkeit der Prognose.

Mit Hilfe des Konfidenzintervalls wird letztlich der Bereich um einen vorher geschätzten Parameter festgesetzt, in dem der wahre Parameter mit einer vorher festgesetzten Wahrscheinlichkeit liegt. Es wird also zusätzlich eine Schätzfunktion benötigt, mit der die wahre Lage relativ genau vorherbestimmt werden muss.

Das Konfidenzintervall selbst definiert sich durch seine Grenzen g_u und g_o und die festgelegte Wahrscheinlichkeit α.

Das Konfidenzintervall für eine Standardnormalverteilung mit bekannter Standardabweichung σ berechnet sich beispielsweise folgendermaßen:

$$c_2 = -c_1 = Z(1-\frac{\alpha}{2})$$

$$g_u = \bar{x} - \frac{c_2 \sigma}{\sqrt{n}} \quad ; \quad g_o = \bar{x} + \frac{c_2 \sigma}{\sqrt{n}}$$

Wobei Z das (1-α/2)-Quantil der Standardnormalverteilung ist, $\bar{x}$ die geschätzte Lage des gesuchten Parameters und n für die Anzahl der Stichproben steht.

6 Die Personalbedarfsprognose im Call Center

6.1 Ausgangsdaten der Personalbedarfsprognose

Vor Beginn der eigentlichen Personaleinsatzplanung steht die Prognose des Mitarbeiterbedarfs. Die Erstellung dieser Prognose erfolgt im Allgemeinen durch die statistische Auswertung vergangener Zeitreihen die von der ACD-Anlage geliefert werden. Die ACD-Anlage dokumentiert alle über sie laufenden Vorgänge und legt die so gewonnenen Daten in einer Datenbank ab, von der aus sie als Berichte und Statistiken geliefert werden können.

Typische Größen, die dabei von einer ACD-Anlage gemessen werden, sind beispielsweise die Eingangszeiten und die Dauer der Anrufe sowie die Zeit, die ein Anrufer in der Warteschlange verbringt. Auch die Zeiträume in denen ein Agent telefoniert oder beispielsweise eine Bildschirmarbeitspause macht, werden vom System registriert und in der Datenbank festgehalten.

Die Zeitreihen, an denen sich die Personalbedarfsprognose letztendlich orientieren muss, bestehen aus Werten, die auf die kleinsten bei der Planung verwendeten Zeiteinheiten aggregiert sind. Typische Intervalle sind hierbei Halb- oder Viertelstunden. Mit kleineren Intervalllängen wie etwa Minuten, würde sich die Genauigkeit der Planung zwar erhöhen, allerdings soll ja am Ende ein Arbeitszeitplan für Mitarbeiter erzeugt werden, die ohnehin schon mit komplizierten Schichtplänen zurechtkommen müssen. Dem durchschnittlichen Angestellten wäre es wahrscheinlich schwer zu vermitteln, dass er beispielsweise um 8 Uhr 23 Arbeitsbeginn hat, dann von 11 Uhr 57 bis 12 Uhr 27 Pause macht und schließlich um 16 Uhr 4 nach Hause gehen darf. So muss sich die Planung in diesem Punkt von der Perfektion verabschieden und dafür im vorgegebenen Rahmen so exakt wie möglich sein.

Wichtige Größen, die von einer ACD-Anlage in jedem Intervall gemessen werden und deren Zeitreihen bei der Personalbedarfsprognose Verwendung finden können, sind:

- **ankommende Anrufe** – Anzahl aller bei der ACD eintreffender Anrufe [ankAnr]
- **mittlere Gesprächsdauer** [mGz]
- **abgenommene Anrufe** – Anzahl erfolgter Bedienvorgänge [abAnr]
- **abgenommener Anrufe im Servicelevel** – Anzahl bedienter Anrufer mit Wartezeiten unter dem Servicelevel-Zeitlimit [abAnrSl]
- **mittlere Klingelzeit** – Zeit, bis der Agent den zugeteilten Anruf abnimmt [mKz]

- **Verzichter** – Anzahl der Anrufer, die während des Wartens auflegen [Verz]
- **Kurzzeitverzichter** – Anzahl der Anrufer, die während des Wartens aber unterhalb des Servicelevel-Zeitlimits auflegen [KuVerz]
- **Besetztfälle** [Bes]
- **mittlere Sperrzeit der Agents** [mSz]
- **angemeldete Bedienplätze** – durchschnittliche Anzahl angemeldeter Agents [Pl]

Alle Durchschnittszeiten werden üblicherweise in Sekunden gemessen.

Das Besetztzeichen bekommen die Kunden zu hören, die zu einem Zeitpunkt anrufen, in dem alle Leitungen des Call Centers belegt sind. Die Anzahl der Leitungen wiederum ergibt sich aus der Summe der angemeldeten Agents zuzüglich der Anzahl der möglichen Wartepositionen.

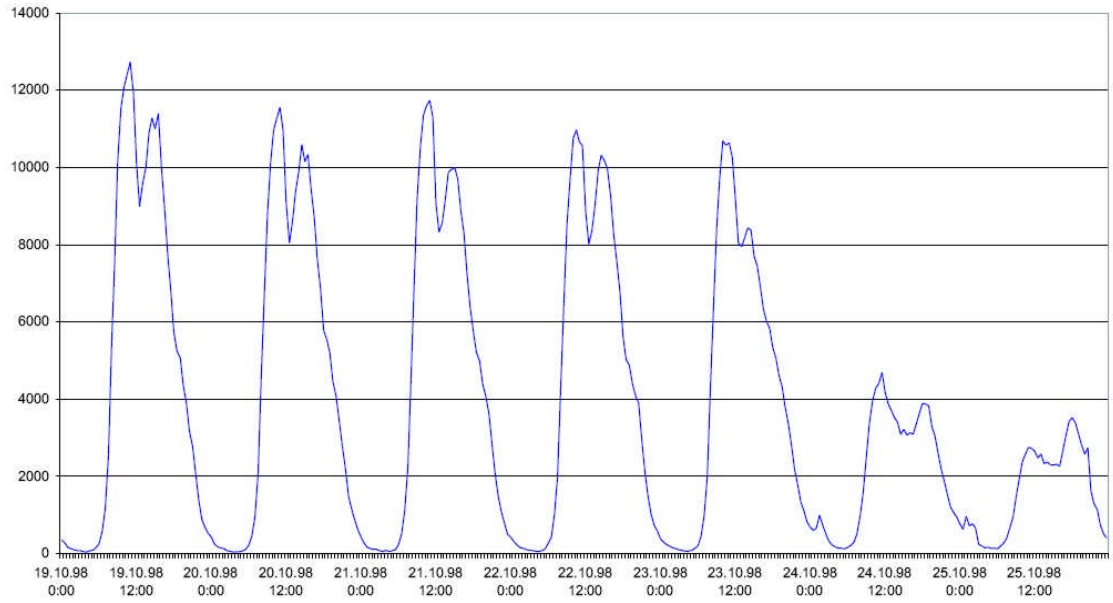
Unter den Sperrzeiten sind zur Vereinfachung des Modells hier gleich zwei Merkmale zusammengefasst. Die Form der Arbeitsunterbrechung, die allgemein unter der Bezeichnung Sperrzeit bekannt ist, umfasst normalerweise nur Zeiträume, in denen sich der Agent kurzfristig gegen die Zuteilung eines neuen Anrufs blockiert. Ein Grund hierfür kann beispielsweise ein Hustenanfall oder das Bedürfnis nach dem Putzen der Nase sein. Daneben hat der Agent aber ab seiner zweiten Arbeitsstunde noch einen gesetzlichen Anspruch auf eine sogenannte Bildschirmarbeitspause (BAP) von fünf Minuten pro Stunde. Diese Unterbrechungszeiten werden im Gegensatz zu den Halbstundenpausen, die ab einer Arbeitszeit von sechs Stunden gesetzlich vorgeschrieben sind, im Normalfall nicht explizit geplant. Da sie für die Personalbedarfsermittlung aber keinesfalls unerheblich sind, sollten ihre mittleren Anteile in die Prognose mit einfließen.

Die aufgelisteten Merkmale genügen im Normalfall zur Ermittlung einer relativ zuverlässigen Prognose des Mitarbeiterbedarfs. Besonders die ankommenden Anrufe und die durchschnittliche Gesprächsdauer sind grundlegende Systemgrößen, die sich in jedem Prognosemodell für Call Center wieder finden. Für den Fall des Best Service Routings werden für den Aufbau eines angemessenen Prognosemodells im späteren Verlauf allerdings noch weitere Größen benötigt, da viele Informationen nur indirekt, über verschiedene Serviceparameter wie beispielsweise die durchschnittliche Wartezeit, herzuleiten sind.

6.2 Die Prognose des Anrufaufkommens

Die Menge ankommender Anrufe [ankAnr] ist die wichtigste Steuergröße jedes IB-Betriebs. Abbildung 4.1 zeigt einen typischen Wochenverlauf des Anrufaufkommens eines IB Call Centers. Zu erkennen ist, dass zwischen den Wochentagen zum Teil erhebliche Unterschiede in Form und Ausschlag der Kurve bestehen können. Besonders an Wochenendtagen ist die Anzahl der Anrufer im Normalfall wesentlich niedriger als an Werktagen. Bei der Prognose für einen bestimmten Wochentag werden deshalb normalerweise auch nur Zeitreihen gleicher Wochentage verwendet.

Soll nun beispielsweise ein Prognosewert für das Anrufaufkommen zwischen 12 Uhr 30 und 12 Uhr 45 eines Montags berechnet werden, so könnte ein sehr einfaches Vorgehen sein, einen Mittelwert über dasselbe Zeitintervall vergangener Montage zu bilden.



Quelle: Helber (2003), S. 5.

Abb. 4.6.1: Wochenverlauf des Anrufaufkommens

Das Anrufaufkommen im Gesamtsystem des Best Service Routings wird einem ähnelnden Verlauf wie in der Abbildung folgen. Der Verlauf des Anrufaufkommens an einem einzelnen Standort muss allerdings keinen anteilig reduzierten aber ansonsten gleichen Verlauf wie das Gesamtaufkommen aufweisen. Eine triviale Lösung, die beispielsweise aussagt, dass ein Call Center im Best Service Routing mit doppelt so vielen Agents im Schnitt auch doppelt so viele Anrufe zugeteilt bekommt, wäre zwar wünschenswert, muss aber vorher im Rahmen einer Studie geklärt werden.

6.3 Die Produktivität der Agents

Die Anrufprognose allein ist natürlich nicht ausreichend, um auf den Mitarbeiterbedarf zu schließen. Eine weitere wichtige Kenngröße vieler Call Center ist die Produktivität [P] der Agents. Diese gibt Auskunft darüber, wie viele Anrufe ein einzelner Agent in einem einzelnen Planungsintervall bewältigen kann.

Bei der Frage, ob diese Produktivität zu berechnen oder lieber festzulegen ist, kann es gelegentlich unterschiedliche Auffassungen geben. Aus Managementsicht ist die Steigerung der Produktivität ein zentrales weil wirtschaftlich relevantes Anliegen. So kann es passieren, dass ein Zielwert vorgegeben wird, nach dem auch geplant werden soll. Es ist hierbei aber wichtig, zwischen strategischen und kurzfristigen Zielen zu unterscheiden. Eine nachhaltige Verbesserung der Produktivität ist nur durch gezielte Trainingsmaßnahmen erreichbar, welche meist auf die Vermittlung von optimierten Gesprächsleitfäden an die Agents abzielen. Mit einer Produktivitätssteigerung kann deshalb nur mittel- bis langfristig gerechnet werden. Unmittelbares Ziel ist aber, die gewünschte Servicequalität zu erreichen. Deshalb ist es immer ratsam, die Produktivität zu berechnen bzw. zu prognostizieren. Ein sich eventuell abzeichnender positiver Trend kann hierbei natürlich in der Prognose berücksichtigt werden. Eine relativ leicht zu implementierende Methode, die auch in kommerziellen Prognosesystemen häufig anzutreffen ist, stellt beispielsweise die exponentielle Glättung zweiter Ordnung.

Die durchschnittliche Gesprächsdauer oder mittlere Gesprächszeit [mGz], ist die einflussreichste Größe, die zur Produktivitätsberechnung herangezogen wird. Sie kann im Tagesverlauf sehr großen Schwankungen unterliegen. Diese können beispielsweise bei einer Telefonauskunft entstehen, weil in der ersten Tageshälfte hauptsächlich Geschäftskunden anrufen, die unter Zeitdruck nur kurze schnelle Informationen suchen. Dagegen überwiegen im Abendbereich die privaten Anrufer, die sich mehr Zeit nehmen können und manchmal auch einfach nur jemanden zum Zuhören suchen.

Die mittlere Gesprächsdauer, wie auch alle weiteren Größen, sollten also, ähnlich wie das Anrufaufkommen, für jedes Planungsintervall separat prognostiziert werden. Die einfachste Form einer verwendbaren Produktivität ergibt sich aus der Formel:

$$P = \frac{t}{mGz}$$

Wobei t für die Zeitdauer eines Planungsintervalls steht. Genauer wird der Wert, wenn die mittlere Sperrzeit vorher von t abgezogen und die mittlere Klingelzeit zur mittleren Gesprächszeit hinzuaddiert wird, denn kein Agent kann ohne Unterbrechungen telefonieren.

Die Anzahl der Anrufe, die ein durchschnittlicher Agent also maximal innerhalb eines Viertelstundenintervalls bewältigen kann, ergibt sich aus der Formel:

$$P = \frac{900 - mSz}{mGz + mKz}$$

Die mittlere Sperrzeit bezieht sich nicht, wie die mittlere Klingelzeit, auf einzelne Anrufe, sondern gibt die Zeitdauer an, die ein im System angemeldeter Agent durchschnittlich nicht gesprächsbereit ist. Bei Verwendung für die Produktivitätsberechnung muss sie also von der Zeitdauer des betrachteten Intervalls abgezogen werden, um so die Zeitdauer zu erhalten, in der ein durchschnittlicher Agent tatsächlich gesprächsbereit ist.

6.4 Bedarfsrechnung mit Erlang-C

Mit den bisher aufgeführten Größen ist es schon möglich, eine ungefähre Abschätzung des erwarteten Bedarfs an Agents [c] aufzustellen. Durch Division der erwarteten Anrufmenge durch die Produktivität, erhält man die Anzahl Agents, die in einem idealisierten Call Center notwendig wäre, um alle Anrufe bewältigen zu können.

$$c = a = \frac{AnkAnr}{P} \qquad \text{oder} \qquad c = a = \frac{AnkAnr * (mGz + mKz)}{900 - mSz}$$

Mit einem idealisierten Call Center ist in diesem Zusammenhang ein Call Center gemeint, dass über einen unbegrenzten Warteraum verfügt. Außerdem wird vorausgesetzt, dass die Anrufer des Call Centers über unbegrenzte Geduld verfügen und deshalb niemals auflegen. Servicelevel-Agreements dürfen in einem solchen Call Center natürlich auch keine Rolle spielen.

Dieselbe Formel, allerdings ohne Berücksichtigung der mittleren Klingel- und Sperrzeiten, wurde schon 1909 von Erlang aufgestellt, um an ihr die minimale Anzahl Agents für ein stabiles System zu ermitteln. Dieser Wert trägt auch die Bezeichnung Arbeitslast [a].

$$a = \frac{AnkAnr * mGz}{t}$$ (für Viertelstundenintervalle gilt: t = 900 Sekunden)

Mit der kompletten Erlang-C Formel ist es darüber hinaus möglich, sich, unter Vorgabe eines gewünschten Servicelevelziels, der dafür benötigten Agentanzahl zu nähern. Die Abbildung 4.2 zeigt die Formel, wobei c für die Anzahl der Agents, a für die Arbeitslast und μ für die Bedienrate ($\mu = 1/mGz$) steht. P[W] ist die Wahrscheinlichkeit überhaupt zu

warten und P[W≤t] die Wahrscheinlichkeit höchstens t Sekunden warten zu müssen. Um die Formel für den schon beschrieben Servicelevel 80/20 anzuwenden, müsste P[W≤t] ≤ 0,8 und t = 20 Sek gesetzt werden.

$$P_1 = P[W] = \frac{\frac{a^c}{c!} \cdot \frac{c}{c-a}}{\left(\sum_{n=0}^{c-1} \frac{a^n}{n!}\right) + \frac{a^c}{c!} \cdot \frac{c}{c-a}}$$

$$P[W \leq t] = 1 - P_1 \cdot e^{-\mu(c-a)\cdot t}$$

Abb. 6.2: Die Erlang-C Formel

Die Wahrscheinlichkeit P[W], dass ein ankommender Anruf alle Bedienstationen belegt vorfindet und demnach warten muss, lässt sich durch die Betrachtung der möglichen Zustandswahrscheinlichkeiten herleiten. Mit Zuständen sind hierbei die Systemzustände gemeint, auf die ein ankommender Anruf treffen kann, also beispielsweise eine ganz bestimmte Anzahl belegter Bedienstationen oder eine ganz bestimmte Anzahl Anrufe auf den Wartepositionen. Alle einzelnen Zustandswahrscheinlichkeiten lassen sich ausgehend von einer Markov-Kette ermitteln. Die Wahrscheinlichkeit, dass ein Anrufer überhaupt warten muss, ergibt sich dann durch das Aufsummieren aller Zustandswahrscheinlichkeiten von dem Zustand an, bei dem alle Bedienstationen belegt sind.

Für eine Herleitung aller Zustandswahrscheinlichkeiten sei an dieser Stelle auf BOSSERT verwiesen (vgl. Bossert (1999), S. 200-201).

Erlang-C berücksichtigt, wie bereits erwähnt, bestimmte Verteilungsformen für die Zwischenankunftszeiten und die Gesprächslängen. Im konkreten Fall sind diese Verteilungsformen Exponentialverteilungen.

Der Term $e^{-(\mu(c-a)*t)}$ in der letzten Erlang-C Formel ist das Integral von t bis Unendlich der gemeinsamen Wahrscheinlichkeitsdichtefunktion der Wartezeiten und liefert die Wahrscheinlichkeit, dass ein wartender Anrufer länger als t Sekunden wartet, in Abhängigkeit von der mittleren Bedienrate μ, der Agentanzahl c und der Arbeitslast a. Allgemein ist eine Wahrscheinlichkeitsdichtefunktion, kurz Dichtefunktion, ein Hilfsmittel, mit dem sich die Wahrscheinlichkeit, dass der Wert einer stetigen Zufallsvariable zwischen zwei reellen Zahlen liegt, bestimmen lässt. Eine stetige

Zufallsvariable ist wiederum eine Funktion, die den Ausgang eines Zufallsexperiments mit reelwertigem Ergebnisraum wiedergibt.

Die Dichtefunktion für die Wartezeiten hat die Formel:

$$W \sim f_w(w) = \mu(c-a)e^{-(\mu(c-a)c)}$$

Wobei W die Zufallsvariable für die Wartezeit ist und letztlich durch die Dichtefunktionen beschrieben wird.

Dichtefunktionen sind unter anderem in der Physik sehr gebräuchlich, um beispielsweise den radioaktiven Zerfall zu beschreiben. Neben der Bezeichnung Dichtefunktion existiert deshalb auch die Bezeichnung Zerfallsfunktion.

Abbildung 4.3 zeigt eine beispielhafte Dichtefunktion für die Verteilung der Wartezeiten.

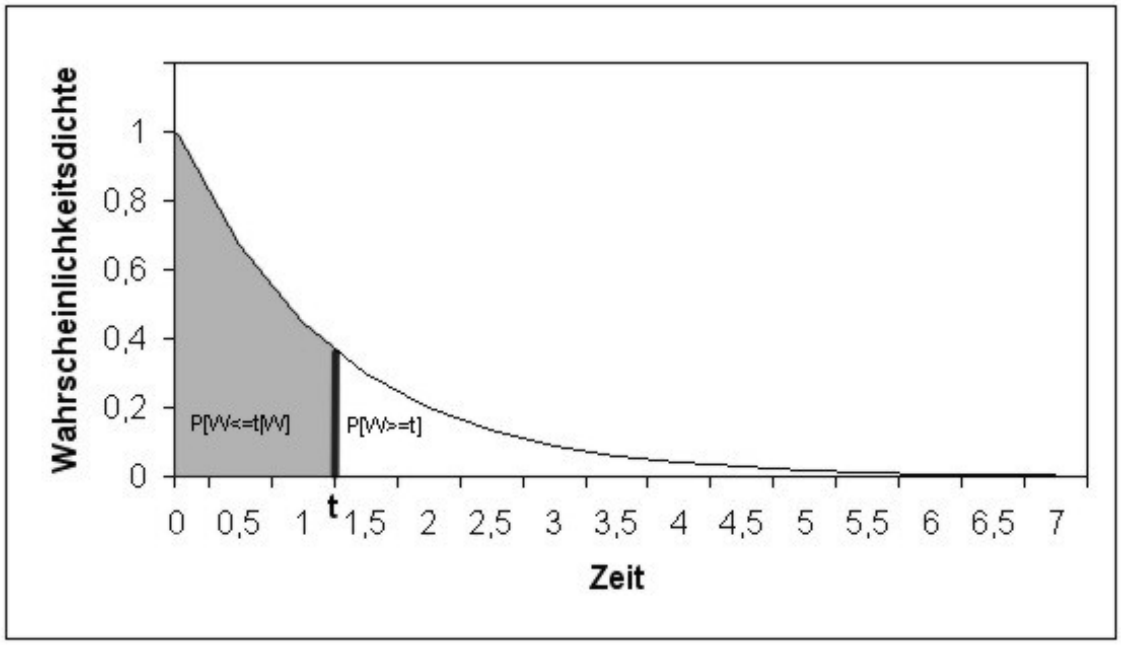


Abb. 6.3: Dichtefunktion der Wartezeitverteilung

Die Verwendung von Exponentialfunktionen ist rein rechnerisch sehr vorteilhaft, weil Integrale einfach zu berechnen sind. In Abbildung 5.3 liefert der Flächeninhalt von Null bis t unter der Kurve, im Verhältnis zum gesamten Flächeninhalt von Null bis unendlich, die Wahrscheinlichkeit, dass ein Anrufer, der sich in der Warteschlange

befindet, weniger oder gleich t Sekunden warten muss. Für unsere Dichtefunktion bestimmt sich dieser Flächenanteil durch:

$$P[W \leq t|W] = 1 - e^{-(\mu(c-a)*t)}$$

Dies ist der Anteil der Wartezeiten zwischen Null und t Sekunden, von der Gesamtanzahl aller Wartenden und wird als die sogenannte Verteilungsfunktion der Exponentialverteilung bezeichnet. Im Umkehrschluss erhält man die Wahrscheinlichkeit, dass ein Wartender länger als t Sekunden warten muss, durch den aus der Erlang-C bekannten Term.

$$P[W \geq t|W] = e^{-(\mu(c-a)*t)}$$

Gewichtet man dies mit der Wahrscheinlichkeit überhaupt zu warten P[W] (siehe Abb. 4.2), so erhält man den Anteil der Anrufer, die bezogen auf die Gesamtanzahl der Anrufer länger als t Sekunden warten müssen. Der restliche Anteil, also die Anrufer, die entweder gar nicht oder weniger als t Sekunden warten müssen, ergibt sich schließlich aus der letzten Erlang-C Formel in Abbildung 6.2.

$$P[W \leq t] = 1 - P[W] * P[W \leq t|W] \qquad (6.1)$$

Um mit Hilfe der Erlang-C Formel den passenden Personalbedarf zu ermitteln, kann c, ausgehend von einem optimistischen Startwert wie der Arbeitslast a, schrittweise um Eins heraufgesetzt werden. Sobald P[W≤t] das erste Mal größer oder gleich der Servicelevelvorgabe wird, ist die beste Lösung gefunden.

Erlang-C basierte Algorithmen finden aufgrund der noch relativ geringen Komplexität häufige Verwendung in kommerzieller Personalplanungssoftware. Formeln, die auf realitätsnäheren Modellen wie (M/M/c/K +M) basieren, sind schon erheblich komplizierter. Meistens sind diese Formeln numerisch problematisch, da sie fast immer komplexe Integrale beinhalten, die nicht nur die Berechnung, sondern auch die Implementierung in eine Anwendung sehr aufwändig machen. Auf eine Herleitung soll an dieser Stelle verzichtet werden. Der interessierte Leser sei dafür auf STOLLETZ verwiesen (vgl. Stolletz (2003), S. 77-79).

6.5 Eine Datenanalyse

Da die Personalbedarfsermittlung auch bei der Lösung des Best Service Routing Problems eine Rolle spielt, soll hier der Versuch einer kleinen Verbesserung des Erlang-C Modells beginnen.

Auf Grundlage einer Datenanalyse soll versucht werden, die beiden vielfach zitierten Schwächen des Erlang-C Modells zumindest näherungsweise zu gewichten, um ausgehend davon, einige Verbesserungen zu entwickeln. Die herausragenden Schwächen waren zum einen die Nichtberücksichtigung des Auflegeverhaltens ungeduldiger Anrufer und die vereinfachende Annahme eines unendlichen Warteraums.

Um das Analyseergebnis vorwegzunehmen, sei schon an dieser Stelle gesagt, dass dem Auflegeverhalten in der Praxis eine ungleich höhere Bedeutung zukommt, als der Berücksichtigung von Besetztfällen, die durch einem endlichen Warteraum verursacht werden. Die Besetztfälle treten im Normalfall erst dann verstärkt auf, wenn ein Call Center entweder einen extrem kleinen Warteraum besitzt oder weit unterhalb seiner Servicelevel-Vorgabe arbeitet. Bei dem Versuch einer Erweiterung des Erlang-C Modells soll also davon ausgegangen werden, dass moderne Call Center mit einer angemessenen Warteraumkapazität ausgestattet sind. Weiterhin wird vorausgesetzt, dass die Personalbedarfsplanung darauf abzielt, einen in der Praxis gebräuchlichen Servicelevel, wie beispielsweise 80/20, zu erreichen. Demzufolge wird es nicht notwendig sein, die Besetztfälle in die später folgenden Betrachtungen mit einzubeziehen, da sie bei korrekter Planung im Prinzip nicht auftreten dürften.

Die Analyse, auf die sich die eben getätigten Aussagen stützen, wurde auf Grundlage von ACD-Berichten eines mittleren IB Call Centers mit ca. 150 Bedienstationen und einem Warteraum von ca. 60 zusätzlichen Telefonleitungen erstellt. Für größere Call Center soll deshalb hier keine Aussage gemacht werden. Es ist aber aufgrund der in Abschnitt 2.2 beschriebenen Skaleneffekte zu erwarten, dass sich, bei steigender Größe, der Einfluss der zufallsbehafteten Größen noch weiter relativiert. Bei gleichem Verhältnis von Bedienstationen zu Wartepositionen, sollten in einem noch größeren Call Center sowohl der Anteil der Besetztfälle als auch der Anteil der Aufleger zurückgehen.

Das untersuchte Call Centers verwendet den wohl am häufigsten eingesetzte Servicelevel 80/20 (siehe Abschnitt 2.3.2). Für abweichende Servicevorgaben wie beispielsweise 90/20 sind die Ergebnisse also mit Vorsicht zu genießen, da auf dieser Basis keine Auswertung stattgefunden hat.

Bei der durchgeführten Analyse, wurden die Berichte von zwei als durchschnittlich anzusehenden Wochen verwendet. Die Intervallwerte auf Viertelstundenbasis wurden zuerst einer Filterung anhand der Servicelevelvorgabe unterzogen. Danach blieben nur die Intervalle übrig, in denen ein durchschnittlicher Servicewert zwischen 80 und 90 Prozent erreicht wurde. Letztendlich blieben von vormals 896 Zeitintervallen (Viertelstundenintervalle von 07:00 Uhr bis 23:00 Uhr an 14 Tagen) gerade einmal 154

übrig, also ca. 17 Prozent. Ein Auszug aus diesen gefilterten Intervallreihen ist in Anhang A zu finden.

Die Untersuchung ergab, dass in diesen als relativ optimal anzusehenden Intervallen gerade einmal 77 Besetztfälle aufgetreten waren. Das entspricht etwa 2 Besetztfällen pro Stunde und ca.0,14 Prozent der gesamten ankommenden Anrufe. Aufgrund dieses Ergebnisses ist davon auszugehen, dass bei einer derartigen Konfiguration keine Berücksichtigung der Besetztfälle notwendig ist, da immer auf ein möglichst genaues Erreichen der Servicelevelvorgabe geplant wird.

Auf der anderen Seite spielten die Aufleger in den betrachteten Intervallen eine wesentlich gewichtigere Rolle. Im Durchschnitt wurden etwa 8,4 Prozent aller ankommenden Anrufe abgebrochen bevor es zur Bedienung durch einen Agent kam. Dieser Wert ist verblüffend hoch, gerade weil das Call Center in den betrachteten Intervallen eigentlich optimale Ergebnisse erzielt hat. Die Berücksichtigung der Aufleger bei der Personalbedarfsermittlung muss also geradezu als notwendig eingestuft werden.

6.6 Erweiterung des Erlang-C Modells

Bei Rekapitulation des in Abschnitt 6.4. dargestellten Wissens über Wahrscheinlichkeitsverteilungen und Exponentialfunktionen, stellt sich die Einbeziehung der ungeduldigen Aufleger, in das Erlang-C Modell, als zumindest näherungsweise lösbare Aufgabe heraus. Die im Folgenden erarbeitete Herleitung erhebt nicht den Anspruch, das theoretische Optimierungsproblem vollständig zu durchdringen, sondern soll die ursprüngliche Erlang-C Formel lediglich um einige Terme erweitern, die sie für einen praktischen Einsatz überhaupt erst konkurrenzfähig und für das Optimierungsmodell des Best Service Routings einsatzfähig machen.

In der Literatur existieren einige Erweiterungen des Erlang-C Modells, wie beispielsweise Palm/Erlang-A. Die Berechnung der meisten dieser Lösungsvarianten ist numerisch aber sehr problematisch, da die Lösungsformeln neben vielfältig verschachtelten Zustandswahrscheinlichkeiten, in den meisten Fällen auch komplexe Integrale enthalten. Weiterhin können endlose Summen auftreten, die wiederum über spezielle Funktionen näherungsweise gelöst werden müssen. Da das Problem ein Gegenstand aktueller Forschung ist, werden wohl auch in Zukunft noch einige neue Berechnungsvarianten für die verschiedenen Warteschlangenmodelle auftauchen.

Die am häufigsten anzutreffenden Formelsammlungen basieren heute auf Warteschlangenmodellen, die sowohl die Ungeduld der Anrufer als auch die endlichen Warteräume berücksichtigt. Der Einfluss der endlichen Warteräume wurde im letzten Abschnitt aber als für die Praxis relativ unbedeutend eingeordnet. Daneben berücksichtigen viele Lösungsansätze auch noch das so genannte Zurückscheuen vor der Warteschlange. Da das Zurückscheuen aber in allen dem Autor bekannten Call Centern keine gemessene Größe ist, wird in dieser Arbeit davon ausgegangen, dass jeder Anrufer entweder:

- nur wartet und auflegt
- nur bedient wird
- wartet und dann bedient wird

Im Hinblick auf die Entwicklung einer relativ überschaubaren Formel und zugunsten einer einfachen Implementierung in einer Anwendung sowie hoher Performance des Prognosesystems, wurde also auf einem Wartemodell mit der Notation (M/M/c/∞ +M) aufgebaut. Eine Literaturliste mit Veröffentlichungen zu unterschiedlichen Modellierungen findet sich bei HELBER (vgl. Helbert (2003), S. 36-37).

Um eine neue Formel zu entwickeln soll an dieser Stelle direkt auf der ursprünglichen Erlang-C Formel aufgesetzt werden, wobei diese unter der Voraussetzung der Nichtexistenz eines Auflegeverhaltens als richtig angesehen wird.

Das Erlang-C Modell geht davon aus, dass jeder Anrufer, der in der Warteschlange landet, auch irgendwann bedient wird, da er ja niemals vor Erhalt des Services auflegt. Da nun aber der Fall mit einbezogen werden soll, bei dem es sehr wohl möglich ist, dass ein Anrufer, auch ohne den Service zu erhalten, auflegen kann, müsste die letzte Erlang-C Formel nun nicht mehr einfach P[W$\leq$t] heißen, sondern P[S$\leq$t], also die Wahrscheinlichkeit der Service mit einer maximalen Wartezeit von t Sekunden zu erhalten.

Um ein allgemeines Auflegeverhalten wartender Anrufer zu berücksichtigen, soll von der gleichen Verteilungsart wie bei den Zwischenankunftszeiten und den Anrufdauern ausgegangen werden. Es soll also im Klartext wieder die komfortable Exponentialverteilung Verwendung finden.

Legt ein wartender Anrufer nach durchschnittlich 20 Sekunden auf, so ist die Auflegerate [v] gleich 1/20 pro Sekunde oder 60/20=3 pro Minute. Dieser Wert wird neben die Zeitvariable t als zusätzlicher Faktor in den Exponenten der Dichtefunktion

erhoben. Die Dichtefunktion für die Verteilung des Auflegeverhaltens hat nun die Formel:

$$f(t) = ve^{-vt}$$

Die dazu gehörige Verteilungsfunktion hat die Formel:

$$F(t) = 1 - e^{-(v*t)} = P[A \leq t \mid \neg S]$$

Da dies ja die Wahrscheinlichkeit liefert, mit der ein nicht bedienter Anrufer bis zum Zeitpunkt t auflegt, könnte auf den ersten Blick vermutet werden, dass nun P[A≤t|¬*S*] einfach zu P[W≤t|W] aus Kapitel 6.4 hinzumultipliziert werden kann, um so bei gegebenem t den Anteil der Aufleger unter den Anrufern zu bestimmen, die zwischen Null und t Sekunden auflegen. Das dies nicht geht, wird klar, wenn vor Augen geführt wird, dass P[A≤t|¬*S*] eine Wahrscheinlichkeit ist, die sich auf eine Situation bezieht, in der Anrufer einfach nur warten, ohne jemals einen Service zu empfangen. Um aber den Anteil der wartenden Anrufer eines realen Call Centers, die zwischen Null und t Sekunden auflegen, zu bestimmen, also P[A≤t|W], muss bewusst werden, dass die wartenden Kunden keines Wegs alle bis zum Zeitpunkt t warten müssen. Der größte Teil von ihnen wird schon vorher von einem Agent bedient. Abbildung 6.4 veranschaulicht die Wartezeiten von Anrufern.

Anrufer 3 stellt in diesem Beispiel den gesuchten Auflegeranteil unter den Wartenden Anrufern dar. P[A≤t|W] = 1/6.

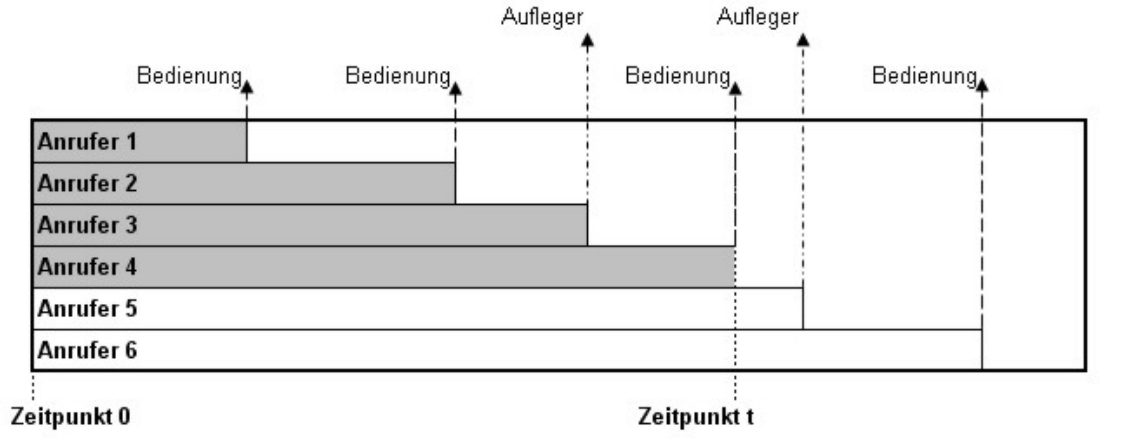


Abb. 6.4: Wartezeiten in der Queue

Die Ereignisse, nach dem Warten den Service zu empfangen und vor Erhalt des Services aufzulegen, schließen sich gegenseitig aus. Wenn schon aufgelegt wurde, kann kein Service mehr erfolgen und umgedreht kann, wenn sich ein Agent dem wartenden Kunden annimmt, dieser nicht mehr zu den Auflegern gezählt werden. Die Wahrscheinlichkeit mit der ein Anrufer innerhalb einer festgesetzten Zeit auflegt, wenn er nicht bedient würde, und die Wahrscheinlichkeit mit der er innerhalb desselben Zeitraumes bedient würde, wenn er nicht auflegt, sind durch die genannten exponentiellen Verteilungsfunktionen berechenbar.

Es sollen nun die Wahrscheinlichkeiten berechnet werden, dass ein wartender Anrufer bis spätestens zum Zeitpunkt t bedient wird und nicht vorher auflegt bzw. bis spätestens zum Zeitpunkt t auflegt und nicht vorher bedient wird. Hierzu modellieren wir die beiden unabhängigen Zufallsvariablen X und Y mit den beschriebenen Exponentialverteilungen des Auflegeverhaltens und den Zeiten bis zur Bedienung durch einen Agent. X steht für die Zeit bis zur Bedienung und Y für die Zeit bis zum Auflegen durch den Kunden.

$$X : f_x(x) \sim \mu(c-a)e^{-\mu(c-a)x}$$

$$Y : f_y(y) \sim ve^{-vy}$$

Um nun beispielsweise die Wahrscheinlichkeit zu berechnen, mit der ein wartender Anrufer bis zum Zeitpunkt t auflegt ohne vorher bedient zu werden, müssen wir zur Bestimmung einer gemeinsamen Verteilung folgendes Doppelintegral über die Dichtefunktionen bilden:

$$P[Y \le X \cap Y \le t] = \int_{y=0}^{t} \int_{x=y}^{\infty} f_y(y) * f_x(x)\, dx\, dy$$

Anschaulich gesprochen ist das äußere Integral die vorhin modellierte Wahrscheinlichkeit, dass ein nicht bedienter Anrufer bis zum Zeitpunkt t auflegen wird. Durch das innere Integral wird zu dieser Wahrscheinlichkeit aber für jeden Zeitpunkt zwischen 0 und t wiederum die Wahrscheinlichkeit hinzumultipliziert, dass der Anrufer später als bis zum Zeitpunkt des Auflegens bedient wird.

$$P[Y \le X \cap Y \le t] = \int_{y=0}^{t} f_y(y) * P[X \ge Y]dy = \int_{y=0}^{t} ve^{-vy} * e^{-\mu(c-a)y}\, dy$$

Wir integrieren mit partieller Integration und erhalten:

$$\int_{y=0}^{t} ve^{-vy} * e^{-\mu(c-a)y}\,dy = -\frac{ve^{-vy}e^{-\mu(c-a)y}}{\mu(c-a)} - \frac{v}{\mu(c-a)}\int_{y=0}^{t} ve^{-vy}e^{-\mu(c-a)y}\,dy$$

$$\int_{y=0}^{t} ve^{-vy} * e^{-\mu(c-a)y}\,dy = \left[-\frac{ve^{-vy}e^{-\mu(c-a)y}}{v+\mu(c-a)}\right]_0^t$$

Der Anteil der Aufleger, die zwischen Null und t Sekunden auflegen, ergibt sich also aus der Formel:

$$P[Y \le X \cap Y \le t] = v\left(\frac{1-e^{-vt}e^{-\mu(c-a)t}}{v+\mu(c-a)}\right)$$

Der Anteil der zwischen Null und t Sekunden bedienten Anrufer ergibt sich analog aus der Formel:

$$P[X \le Y \cap X \le t] = \mu(c-a)\left(\frac{1-e^{-vt}e^{-\mu(c-a)t}}{v+\mu(c-a)}\right)$$

Abbildung 6.5 zeigt die Wahrscheinlichkeit, dass ein wartender Anrufer in unter t Sekunden bedient wird *P[X≤Y∩X≤t]:*

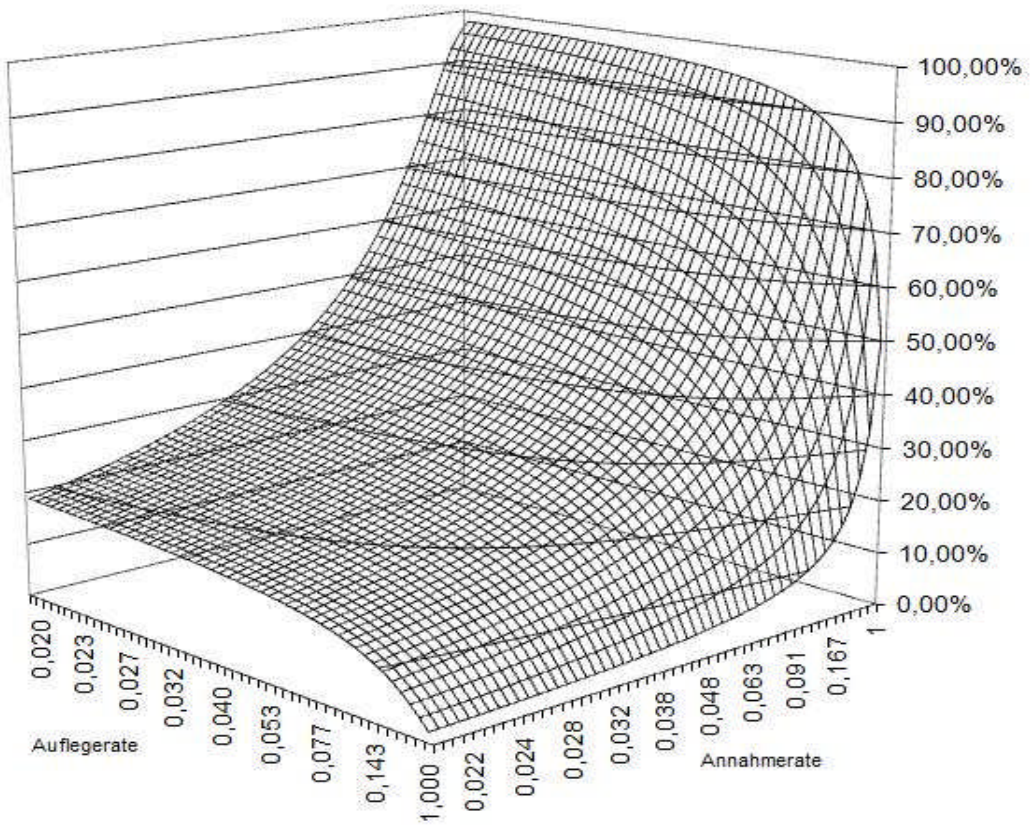


Abb. 6.5: gemeinsame Verteilungsfunktion

Die endgültige Formel für die Berechnung des Servicelevels setzt sich nun wie folgt zusammen:

$$P[W \le t] = 1 - P[W] * (1 - P[Y \le X \cap Y \le t] - P[X \le Y \cap X \le t]) \quad (6.2)$$

(1-(P[Y≤X∩Y≤t]-P[X≤Y∩X≤t]) ist hierbei der Anteil der wartenden Anrufer, die nicht unterhalb des Zeitlimits t bedient werden oder auflegen. Multipliziert mit der Wahrscheinlichkeit überhaupt zu warten P[W], ergibt sich der Anteil der Calls mit schlechtem Service von der Gesamtheit aller ankommenden Anrufe. P[W] wurde komplett aus der Erlang-C Formel übernommen.

Die Formel gibt nun einen Servicelevel an, der sich immer etwas über der Erlang-C Vorhersage bewegt. Da Erlang-C den realen Servicelevel fast immer unterschätzt, ist dies ein gewünschter Effekt. In ersten Tests an realen Datenreihen, ergab sich, im Vergleich zu Erlang-C, eine um 18 Prozent geringere absolute Abweichung von den wirklich erreichten Serviceleveln. Dabei wurden allerdings nur die Intervalle ausgewertet, für die Erlang-C einen Servicelevel von mindesten 50 Prozent vorhergesagt hatte. Bezogen auf die kompletten Datenreihen war die Verbesserung sogar noch deutlich größer, da Erlang-C den Mitarbeiterbedarf für sehr niedrige Servicelevel kaum noch realitätsnah abschätzt. Der Grund dafür ist, dass bei längeren tolerierten Wartezeiten der Anteil der Aufleger naturgemäß größer wird und dies von Erlang-C überhaupt nicht berücksichtigt wird.

Da sich der Servicelevel aber in der Praxis oft nicht aus dem einfachen Verhältnis von abgenommenen Anrufen innerhalb der Servicevorgabe zu insgesamt ankommenden Anrufen ergibt, muss an dieser Stelle noch eine alternative Formel bereitgestellt werden. Bei der häufigsten Berechnungsvariante des Servicelevels wird die Menge der ankommenden Anrufe von vornherein um den Anteil der Anrufer verringert, die innerhalb der Servicevorgabe auflegen. In der Erlang-C Formel findet dies leider überhaupt keine Berücksichtigung, aber um der Realität mit der neuen Formel wenigstens halbwegs gerecht zu werden, soll dieser Fakt hier nicht unterschlagen werden.

In Abbildung 6.6 ist der Anteil der sogenannten Kurzzeitverzichter durch das rechte unteren Rechteck dargestellt. In dieser im Vergleich zur bisherigen Darstellung alternativen Symbolik steht S für das Anrufereignis des Serviceerhalts also das bedient werden und A für das Ereignis des Auflegens. Das mit dem Pipezeichen abgetrennte W steht für die voraus geltende Bedingung des Wartens. In der bisherigen Darstellung wurde das Warten meist vorausgesetzt, nun soll dies explizit herausgestellt werden. P[S≤t|W] ist hier beispielsweise synonym zu *P[X≤Y∩X≤t]* zu sehen.

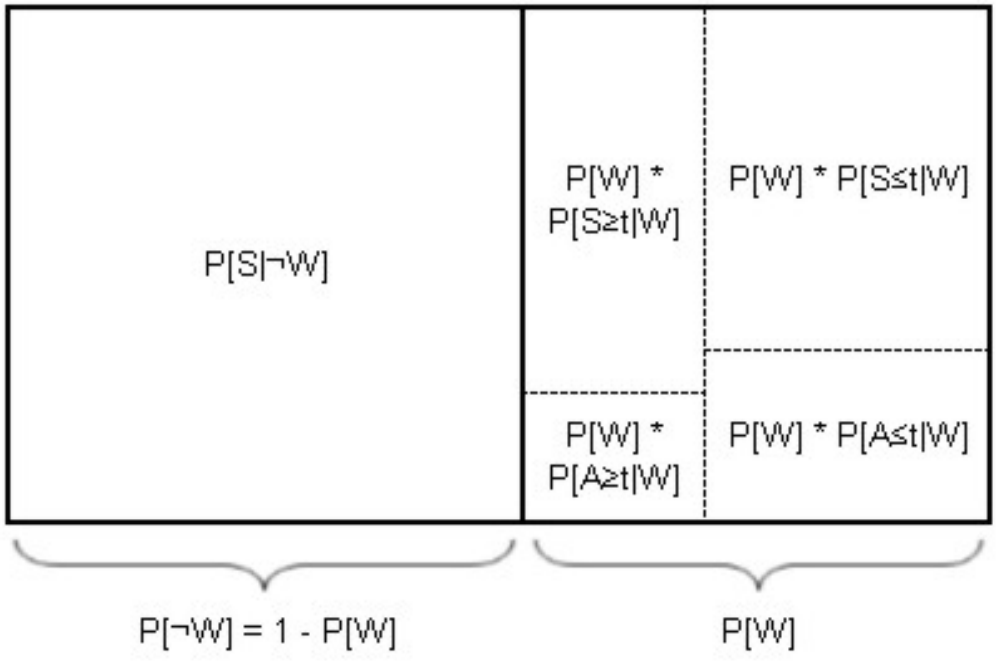


Abb. 6.6: Aufteilung der Anrufergebnisse

Um den Servicelevel unter den beschriebenen Voraussetzungen abschätzen zu können, lautet die Formel jetzt:

$$P[W \le t] = \frac{(1 - P[W]) + P[W] * P[X \le Y \cap X \le t]}{1 - P[W] * P[Y \le X \cap Y \le t]} \qquad (6.3)$$

Im Zähler steht hierbei der Anteil der Anrufer, die den Service sofort erhalten, zuzüglich dem Anteil der Anrufer, die weniger als t Sekunden auf den Service warten. Im Nenner findet sich der Gesamtanteil aller ankommenden Anrufe abzüglich der Anrufer, die vor Erreichen der Servicelevelschwelle auflegen. Der Servicelevel, der sich aus dieser Formel ergibt, liegt nun zwischen denen von Formel (4.1) und (4.2) vorhergesagten Werten.

Es muss an dieser Stelle angemerkt werden, dass sich eigentlich auch die Wartewahrscheinlichkeit P[W] ändern muss, sobald die Aufleger berücksichtigt werden. Aus rein logischen Überlegungen müsste P[W] etwas kleiner werden, da die Warteschlange nun neben der Bedienungsannahme noch durch das zusätzlichen Auflegeereignisse abgebaut wird. Der Einfluss der Aufleger auf die Wartewahrscheinlichkeit kann im Vergleich zu seinem Einfluss auf die allgemeine Wartezeit aber nur sehr gering sein. Aufleger machen in diesem Modell immer nur einen kleinen Teil der Wartenden aus. In jedem Fall wirkt sich die Einbeziehung der Aufleger aber auch so schon positiv auf die abgeschätzten Servicelevel aus, da ein

Aufleger die Wahrscheinlichkeit der anderen wartenden Anrufer erhöht, noch innerhalb der Servicelevelvorgabe bedient zu werden. Wird nun P[W] noch zusätzlich verringert, ist mit einem noch höheren Schätzwert für den Servicelevel zu rechnen. Eine genauere Bestimmung von P[W] soll an dieser Stelle nicht erfolgen. Die bis hierhin entwickelte Formel ist zwar nicht perfekt aber liefert schon deutlich bessere Ergebnisse als Erlang-C. Der von ihr gelieferte Schätzwert, ist trotzdem noch als leicht pessimistisch einzuordnen. Dieser Umstand ist allerdings weniger schlimm, denn eine leichte Überschätzung des Mitarbeiterbedarfs ist immer noch besser als das Gegenteil, wenn davon ausgegangen wird, dass eine Unterschreitung des Servicelevels in der Praxis wesentlich höhere Kosten nach sich zieht, als eine Überschreitung.

6.7 Ein Optimierungsmodell für Vertriebs-Call-Center

Die im letzten Abschnitt erarbeitete Formel bezieht sich in erster Linie auf ein Call Center, das einer Servicelevelvorgabe folgen muss. In diesem Abschnitt soll eine grundsätzliche Formel der Bedarfsoptimierung eines Vertriebs-Call-Centers aufgebaut werden, um im weiteren Verlauf auch dafür ein Optimierungsmodell für den Best Service Routing Fall konstruieren zu können.

Vertriebs-Call-Center produzieren Umsatz über den telefonischen Verkauf von Produkten oder Dienstleistungen. Im Durchschnitt trägt also jeder getätigte Anruf direkt zum Unternehmenserfolg bei. Über die Gewinnfunktion eines Vertriebs-Call-Centers soll der Unternehmensgewinn maximiert werden. Neben den Umsätzen muss die Gewinnfunktion natürlich die Kosten, die hauptsächlich durch den Personaleinsatz entstehen, berücksichtigen. Die grundsätzliche Gewinnfunktion eines reinen Vertriebs-Call-Center muss also wie folgt aussehen:

$$G = AUC * AngAnr - c * L$$

AUC ist hierbei der durchschnittliche Umsatz pro Anruf und L der durchschnittliche Lohn eines Agents pro Zeiteinheit.

Vernachlässigen wir wieder die Besetztfälle, so ergibt sich die erwartete Anzahl angenommener Anrufe aus den ankommenden Anrufen abzüglich der erwarteten Aufleger.

$$AngAnr = AnkAnr * (1 - P[A])$$

Der Anteil der Aufleger errechnet sich aus der Wahrscheinlichkeit zu warten multipliziert mit der Wahrscheinlichkeit als Wartender aufzulegen. Unter Verwendung

der im letzten Abschnitt erarbeiteten Formel für die Wahrscheinlichkeit des Auflegens bis zum Zeitpunkt t, ergibt sich folgende Formel:

$$P[A] = P[W] * P[Y \leq X]$$

$$P[Y \leq X] = \int_{y=0}^{\infty} \int_{x=y}^{\infty} f_x(x) f_y(y) \, dx \, dy$$

Y wird diesmal über den kompletten Zeitraum integriert, da diesmal keine Servicelevelbeschränkungen existieren.

$$\int_{y=0}^{\infty} ve^{-vy} * e^{-\mu(c-a)y} dy = \left[-\frac{ve^{-vy} e^{-\mu(c-a)y}}{v + \mu(c-a)} \right]_0^{\infty}$$

$$P[Y \leq X] = \frac{v}{v + \mu(c-a)}$$

Die vollständige zu maximierende Gewinnfunktion hat also die Formel:

$$G = AUC * AnkAnr(1 - (P[W] * \frac{v}{v + \mu(c-a)})) - c * L \qquad (6.4)$$

Es gilt wieder:

$$P[W] = \frac{\frac{a^c}{c!} * \frac{c}{c-a}}{\left(\sum_{n=0}^{c-1} \frac{a^n}{n!} \right) + \frac{a^c}{c!} * \frac{c}{c-a}}$$

und

$$a = \frac{AnkAnr * mGz}{t} \quad \text{oder} \quad a = \frac{AnkAnr * (mGz + mKz)}{T - mSz}$$

Alle Parameter der Gewinnfunktion 6.4, bis auf die Anzahl der Mitarbeiter, sind vorgegeben oder können durch Zeitreihenanalyse über historische ACD-Messwerte prognostiziert werden. Die Suche nach dem Gewinnmaximum könnte also, wie schon im Abschnitt 6.4 für Erlang-C beschrieben, durch einfache Iteration über alle möglichen ganzzahligen Werte von c innerhalb eines realistischen Bereichs vorgenommen werden. Die untere Grenze dieses Bereichs könnte durch den optimistischen Schätzwert c_u = a vorgegeben werden. Falls eine obere Grenze nicht durch beispielsweise Platzbegrenzungen vorgegeben ist kann c auch solange erhöht

werden, bis der Funktionswert der Gewinnfunktion zum ersten Mal wieder kleiner als sein Vorgängerwert wird.

Die Möglichkeit über die Ableitung der Gewinnfunktion das Optimum oder die Optima zu bestimmen, ist vor allen Dingen aufgrund der Verschachtelungen, die bei der Berechnung der Wartewahrscheinlichkeit P[W] auftreten, relativ komplex und soll hier nicht weiter erörtert werden.

6.8 Probleme der Bedarfszahlen

Ein wichtiger Fakt, der selbst von einigen, dem Autor bekannten, kommerziellen Softwarelösungen ignoriert wird, sind die in Abschnitt 2.5.1 erläuterten Sperrzeiten. Die auf Grundlage wahrscheinlichkeitstheoretischer Überlegungen aufgestellten Formeln wie beispielsweise Erlang-C, ermitteln einen Bedarf an einsatzbereiten Bedienplätzen. Nun sind aber in einem realen Call Center keineswegs immer alle angemeldeten Bedienplätze auch einsatzbereit. Der Grund dafür sind die Sperrzeiten. Einige Anwendungen, benutzen eine durch Erlang-C ermittelte Tagesbedarfskurve und verplanen die realen Mitarbeiter genau bis zur Obergrenze der Kurve, als wären sie allzeit bereite Roboter. Mit etwas Glück steht wenigstens noch ein Ausgleichsfaktor zur Einbeziehung eines erwarteten Krankenstandes zur Verfügung, aber mehr hat der Anwender im Normalfall nicht zu erwarten. Dabei ist die Sperrzeit ein wichtiger Einflussfaktor, der unter ungünstigen Umständen sehr starke Einbrüche in der Servicequalität verursachen kann. Man stelle sich nur einmal vor, dass sich mehrere Mitarbeiter verabreden, gemeinsam eine Bildschirmarbeitspause zu machen. Diese Gleichzeitigkeit wird zwar in den meisten Call Centern durch gezielte Steuerung der Mitarbeiter unterbunden, jedoch unterliegen die durchschnittlichen Sperrzeiten im Tagesverlauf trotzdem erheblichen Schwankungen. Wird in einem Call Center nicht rund um die Uhr gearbeitet, so ist die Sperrzeit besonders zu Beginn und zum Ende der Produktivperiode doch erheblich niedriger als im Tagesdurchschnitt. Dies ist auf die übliche Bestimmung zurückzuführen, die den Mitarbeitern die früheste BAP nach einer Stunde ihrer Arbeitszeit und die späteste eine Stunde vor Ende ihrer eingeplanten Arbeitszeit erlauben. Weiterhin sind die meisten Mitarbeiter bestrebt, ihre Kurzpausen stärker in den Mittagsbereich zu verschieben, was wiederum eine Häufung der Sperrzeiten in diesem Tagesbereich verursacht.

Die so entstehenden Produktionsstundenlöcher können also, bei Nichtbeachtung während der Planerstellung, durchaus zu schlechten Ergebnissen im operativen Betrieb führen. Die durchschnittlichen Sperrzeiten sowie der erwartete Krankenstand sollten

deshalb ebenfalls ermittelt bzw. prognostiziert werden und in die Bedarfsplanung einfließen. Eine Möglichkeit dies zu realisieren, ist die Verwendung der im letzten Abschnitt erstellten Formeln wobei sich die Arbeitslast a nicht wie in der Erlang-C Formel aus *a = AnkAnr * mGz / t* berechnet sondern wie am Anfang von Abschnitt 6.4. beschrieben aus *a = AnkAnr * (mGz + mKz) /(900 Sek – mSz).* Bei Einbeziehung von Krankheitsprognosen müsste der gesamte Term rechts neben dem Gleichheitszeichen zusätzlich durch den erwarteten Anteil an nicht Kranken geteilt werden.

$$a = \frac{AnkAnr * (mGz + mKz)}{(900\,Sek - mSz) * P[gesund]}$$

7 Best Service Routing

7.1 Routingvarianten

Bei großen IB-Projekten, wie beispielsweise einer Telefonauskunft oder einer Versandannahme, sind aufgrund des hohen Anrufaufkommens oft mehrere Call Center von häufig auch unterschiedlichen Vertriebspartnern im Einsatz. Eine übliche Strategie, die Anrufe zwischen den einzelnen Standorten zu verteilen, ist eine Unterteilung des Gesamtgebietes, aus dem die Anrufe eingehen können. Jedes Call Center bekommt eines der Teilgebiete zugewiesen, das daraufhin als sein Routinggebiet bezeichnet wird. Ein Routinggebiet kann dabei beispielsweise ein Bundesland sein. Ein eingehender Anruf wird im Normalfall durch seine Ortsvorwahl identifiziert und einem Gebiet zugeordnet. Die ACD-Anlagen der beteiligten Call Center bekommen letztendlich alle Anrufe zugeleitet, die aus dem jeweilig zugeteilten Routinggebiet beim übergeordneten Anrufverteiler eingehen. Jedes Call Center hat bei dieser Routingvariante also in gewisser Weise ein regionales Monopol für die betreute Rufnummer.

Wie schon in Abschnitt 2.2 beschrieben, kann es bei dieser Art des Routings vorkommen, dass Anrufer das Besetztzeichen erhalten, obwohl an einem anderen Standort noch unbeschäftigte Agents sitzen. Eine Balancierungsstrategie, um derartige Kapazitäten zumindest teilweise auszuschöpfen und die Call Center virtuell zusammenarbeiten zu lassen, stellt der sogenannte Overflow dar. Bei diesem Überlaufsystem würde ein Anruf, der an einem Standort geblockt wird, auf den jeweils nächsten Standort umgeleitet.

Best Service Routing (BSR) geht sogar noch einen Schritt weiter. Beim reinen BSR muss es überhaupt keine festen Routinggebiete geben. Alle Anrufe der Servicenummer landen in einem einzigen Pool, aus welchem alle beteiligten Call Center beliefert werden. In einer Situation mit mehr ankommenden Anrufen, als freien Agenten, erfolgt die zentrale Verteilung der Anrufe auf die einzelnen ACD-Anlagen nach einer festen Systematik, die sich nicht mehr am Ursprungsort des Anrufs orientiert, sondern an der sogenannten Expected Wait Time (EWT). Dieser Wert wird bei Eintreffen eines Anrufs für jede ACD-Anlage gemessen. Die Höhe der EWT ist üblicherweise von der erwarteten Zeit abhängig, die ein an die ACD geleiteter Anrufer warten müsste, bis er mit einem Agent verbunden wird. Die ACD, die bei Eintreffen eines Anrufs die niedrigste EWT zurückliefert, bekommt diesen Anruf zugeteilt. Die genaue Berechnung der EWT ist dabei abhängig von den jeweiligen Einstellungen und wird von der Firma AVAYA auch nicht konkret bekannt gegeben. Für die weitere Arbeit soll deshalb vereinfacht davon ausgegangen werden, dass sich die EWT aus der Menge der wartenden Anrufer [WA] in der jeweiligen Queue am Standort s inklusive dem neu

eintreffenden Anruf, multipliziert mit der aktuellen durchschnittlichen Wartezeit in der Queue, ergibt.

Für $WA_s<(K_s-c_s)$ gilt:

$$EWT_s = (WA_s + 1) * AWT$$

für $WA_s=K_s-c_s$ gilt:

$$EWT_s = \infty$$

Ein BSR-System ist somit darauf ausgerichtet, dem Anrufer die geringste mögliche Wartezeit zuzumuten. Mit diesem Ziel werden alle Warteschlangen in etwa gleich stark belastet.

Besteht bei Eintreffen eines Anrufs eine Situation mit einem Überschuss an Agents, was bedeutet, dass mindestens zwei Agents im Gesamtsystem frei sind, so kann die Verteilung durch unterschiedliche Strategien realisiert werden. Beispielsweise könnte der Agent den Anruf erhalten, der am längsten unbeschäftigt wartet oder es könnte der Standort mit der geringsten durchschnittlichen Bearbeitungsdauer beliefert werden. Das eigentliche Best Service Routing greift also nur in Situationen mit Anrufüberschuss. In allen anderen Fällen können beliebige Verteilungsstrategien angewendet werden. Es ist somit auch eine Kombination aus Routinggebiet und BSR denkbar.

Wie bereits in Abschnitt 2.2 erwähnt, soll mit Hilfe von Best Service Routing erreicht werden, dass mehrere Call Center derart zusammenarbeiten, als wären sie ein einziges großes Call Center, dessen Bedien- und Warteplatzkapazität, der Summe der Einzelteile entspricht. Mit der bisher beschriebenen Variante des Best Service Routings ist dies allerdings noch nicht vollständig zu erwarten, da jedes der Call Center immer noch eine eigene Queue besitzt. Ist ein Anruf erst einmal an eine der Queues weitergeleitet worden, so verbleibt er dort, bis er von einem Agent des zugehörigen Call Centers abgenommen wurde, oder bis der Anrufer auflegt. Obwohl die beschriebene Berechnung der EWT darauf abzielt, den Anruf an die Queue zu leiten, in der am ehesten mit einer schnellen Abnahme gerechnet wird, muss das am Ende nicht unbedingt stimmen. Auch die EWT ist kein exakter Wert, sondern eine Prognosegröße, deren Wert sich fortlaufend ändert. Es ist also möglich, dass ein Anruf an eine Queue geleitet wird, die letztlich doch langsamer abgearbeitet wird als erwartet. Dies geschieht beispielsweise wenn während des Wartevorgangs in den anderen Call Centern zufallsbedingt mehr wartende Anrufer auflegen oder plötzlich mehr Agenten frei werden, als zum Zeitpunkt der Routings geschätzt wurde. Für diese Fälle existieren weiter Strategien, die es erlauben, beispielsweise besonders lange wartende Anrufe aus

bestimmten Queues auszuwählen, um diese daraufhin in eine der anderen Queues einzuordnen. Auch dieser Prozess der Auswahl und Neuzuordnung basiert wiederum auf bestimmten Schätzverfahren, die nicht hundertprozentig genau sein können, da beispielsweise nie vorhersehbar ist, wann einzelne Bedienvorgänge abgeschlossen sind und Agents frei werden. Eine derartige Organisation mehrerer Queues wird deshalb nie ganz erreichen, dass sich die einzelnen Queues zusammen wie eine einzige Warteschlange verhalten. Mit zunehmender Verfeinerung der Verfahren ist aber auch eine zunehmende Annäherung an dieses Ziel möglich.

7.2 Das Modell des Best Service Routings

Im Bezug auf das zu modellierende Prognosesystem für das BSR, muss grundlegend von einem Wartesystem ausgegangen werden, dass aus mehreren Untereinheiten besteht. Diese stehen durch ihre gemeinsame Inputquelle zueinander in Beziehung und beeinflussen sich gegenseitig durch ihre spezifischen Kapazitäten, Servicewerte und Bearbeitungsraten.

Abbildung 5.1 verdeutlicht ein solches System an einem Beispiel mit drei beteiligten Standorten.

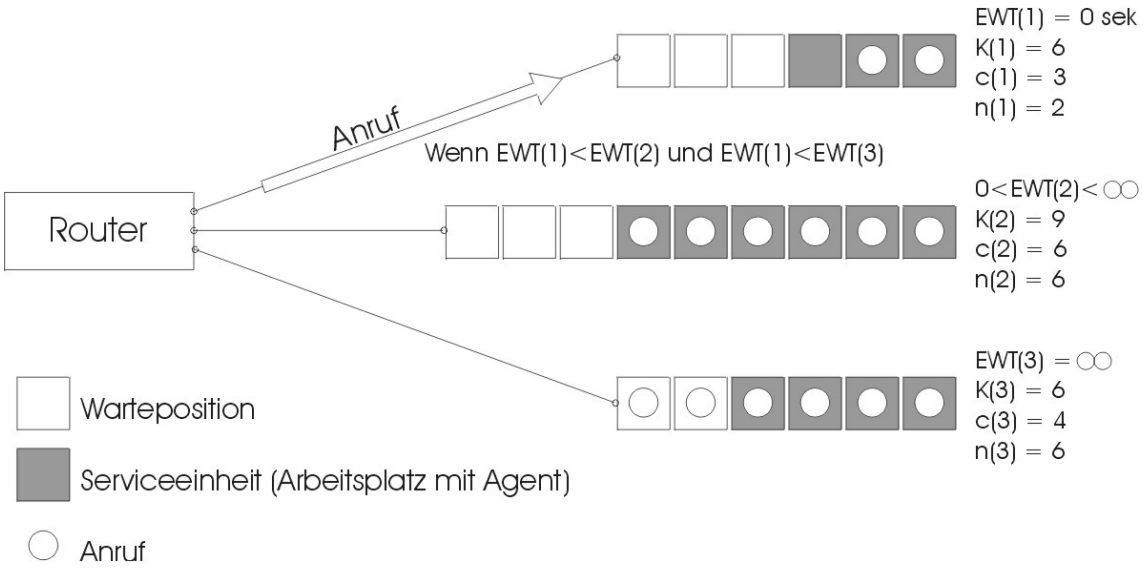


Abb. 7.1: Modell des Best Service Routings

Ein Prognoseverfahren, dass an einem bestimmten Standort eingesetztes wird, muss nun Voraussagen über den Zustand des Gesamtsystems treffen und daraus ableiten können, wie die Auswirkungen auf den eigenen Standort sein werden.

Die Informationen die dem einzelnen Standort dazu zur Verfügung stehen beschränken sich dabei auf die Messwerte, welche durch die eigene ACD-Anlage bereitgestellt werden und die Werte, die am zentralen Verteilerknoten gemessen werden. Es soll davon ausgegangen werden, dass zumindest die zentral gemessenen Werte des gesamten Anrufaufkommens allen Beteiligten zur Verfügung steht. Wäre dies nicht der Fall, würden für keinen beteiligten Standort genügend Informationen bereitstehen, um seinen eigenen Anteil am Gesamtsystem abschätzen zu können und es würde eine unverzichtbare Planungs- bzw. Prognosegrundlage fehlen. Die zentral gemessenen Besetztfälle sollen wieder ignoriert werden, da sie unter Optimalbedingungen in diesem Modell nicht erwartet werden.

Die grundlegenden Parameter, neben den erwarteten ankommenden Anrufen im Gesamtsystem, sind für eine nachträgliche Bestimmung der optimalen Mitarbeitereinsätze im Fall des Dienstleistungs-Call-Center in der Formel 6.3 in Abschnitt 6.6 zu finden und für den Fall des Vertriebs-Call-Centers in Formel 6.4 in Abschnitt 6.7.

Es ist aber notwendig, das mathematisches Grundmodell für einzeln arbeitende Call Center nun für den BSR Fall zu erweitern. Es soll ermöglicht werden aus den vorhandenen Informationen eine Ableitung der fehlenden Informationen vorzunehmen. Nur dann können die richtigen Schlüsse für die Situation am eigenen Standort gezogen werden.

Eine Vereinfachung des Gesamtsystems, durch beispielsweise eine Unterteilung in nur zwei Einheiten, wie etwa eigener Standort und zusammengefasster Rest, kann dabei die Komplexität des erforderlichen Prognoseprozesses deutlich verringern. Dafür ist es allerdings wiederum notwendig, den Effekt des BSR genau zu quantifizieren, um zu entscheiden, ob mehrere Call Center im BSR-System überhaupt als ein einziges darstellbar sind oder ob die Komplexität des Gesamtsystems dies eher nicht zulässt.

7.3 Fragestellungen

Einige Fragen, die im Folgenden im Rahmen einer Simulation geklärt werden sollen sind:

1. Welche Rolle spielt die Größe eines Call Centers im BSR, im Bezug auf die erhaltene Anrufmenge und die erreichten Servicewerte bzw. Gewinne? Bekommt ein doppelt so großes Call Center auch doppelt so viele Anrufe?

2. In welcher Größenordnung ist der Gesamteffekt des BSR im Vergleich zum Standardrouting anzusetzen? Wie groß müsste ein einzelnes Call Center sein, um ähnliche Ergebnisse wie der BSR-Verbund zu liefern?

3. BSR soll die Servicequalität zwischen den Call Centern angleichen und insgesamt erhöhen. Hat ein kleineres Center Nachteile beim Versuch die geforderten Servicelevel zu erreichen oder erreicht es mit beliebiger Besetzungsstärke immer ähnliche Servicewerte wie die anderen?

8 Simulation des Best Service Routing

8.1 Wahl des Simulationsart

Bei der Untersuchung von Systemen, die sich durch hohe Komplexität auszeichnen und deren mathematische Zusammenhänge durch theoretische Überlegungen nur sehr ungenau oder nur unter erheblichem Aufwand herzuleiten sind, können Simulationen oft hilfreiche Informationen liefern. In dynamischen Systemen wie beispielsweise bei dem im letzten Abschnitt skizzierten BSR-Modell, kann durch die Simulation eines Modells auf bestimmte Systemeigenschaften und teilweise auch auf ein allgemeines Systemverhalten geschlossen werden.

Die Wahl einer geeigneten Simulationsmethode hängt dabei stark von der Art des untersuchten Sachverhalts oder Systems ab.

Zur Wiederholung sei an dieser Stelle noch einmal erwähnt, dass alle Warteschlangensysteme, also auch das System des BSR, von stochastischen Prozessen geprägt sind. Weiterhin zeichnen sie sich durch die ständige Wiederholung eines relativ einfachen Bedienvorgangs aus. Im Falle von Call Centern ist dieser Bedienvorgang die Bearbeitung eingehender Anrufe. Die dabei angestrebten Ziele sind, abhängig von der jeweiligen Art des Call Centers, entweder die Maximierung des Gewinns oder die Minimierung der Kosten. Zu diesem Zweck müssen in erster Linie bestimmte durchschnittliche Servicewerte erreicht oder bestimmte Umsätze erwirtschaftet werden. Weitere Leistungsgrößen, wie durchschnittliche Wartezeiten der Anrufer oder die durchschnittlichen Bereitzeiten der Agents, haben ebenfalls einen Einfluss auf Gewinne und Kosten und müssen mitbetrachtet werden.

Derartige Systeme, die sich durch die häufige Wiederholung eines Prozesses darstellen lassen, der zwar beschreibbar, aber im Einzelnen, aufgrund von Zufallseinflüssen, nicht vorhersagbar ist, können mit Hilfe der Monte Carlo Methode simuliert werden. Monte Carlo Simulationsmodelle haben bestimmte Eingangs- und bestimmte Ausgangsobjekte bzw. -variablen. Bei der Monte Carlo Simulation werden so genannte Pseudozufallszahlen generiert. Diese sollen zufällige Ausprägungen realer Merkmale simulieren und müssen deshalb auch ähnlichen Häufigkeits- und Größenverteilungen folgen. Durch häufiges Wiederholen eines Zufallsexperiments entstehen Ergebnisreihen, deren Durchschnittswerte und Verteilungsmuster das eigentliche Simulationsergebnis darstellen. Mit ihnen können Erwartungswerte für reale Sachverhalte anhand des Simulationsmodells bestimmt werden.

8.1.1 Vorüberlegungen zum Simulationsmodell

Bei der Simulation realer Vorgänge ist es wichtig, das untersuchte Problem auf relevante Merkmale zu reduzieren. Ein Simulationsmodell sollte daher nur eine überschaubare Menge an Merkmalen enthalten, die wenn möglich auch voneinander abgegrenzt beobachtet werden müssen. Eine gleichzeitige Variation mehrerer Merkmale lässt schon bei geringer Komplexität des Systems kaum noch Rückschlüsse auf den jeweils ausschlaggebenden Einflussfaktor zu.

Reale Merkmale, die im hier untersuchten BSR-Modell deshalb vollständig unbeachtet bleiben sollen, sind beispielsweise die Sperrzeiten, die Klingelzeiten und Extremfälle wie zurückscheuende Anrufer oder außerordentlich lange Anrufe. Die Begrenzung der Warteräume soll zwar grundsätzlich im Modell mit berücksichtigt werden, es sollen aber, aufgrund der unendlichen Anzahl denkbarer Variationen, nur die Fälle mit jeweils gleichem Verhältnis zwischen Bedienplätzen und Warteplätzen sowie mit unendlich großen Warteräumen betrachtet werden. Weiterhin sollen alle Modell Call Center dieselben durchschnittlichen Bearbeitungszeiten aufweisen, um einen Vergleich nur anhand der Größe durchführen zu können.

8.1.2 Objekte und Variablen des Simulationsmodells

Grundlegende Objekte des Modells sind Anrufe und Call Center. Jedes Call Center besitzt als weitere Objekte eine Queue mit einer bestimmten Anzahl Wartepositionen und eine bestimmte Anzahl gleicher Bedienplätze.

Der Anruf ist das Eingangsobjekt des BSR-Zufallsexperiments und die Ausgangsobjekte sind Ergebnisse der Anrufbearbeitung bzw. Nichtbearbeitung sowie Messgrößen zur Auslastung der Agents. Jeder Anruf hat bestimmte Merkmale, deren Ausprägung bei Betrachtung eines einzelnen Anrufs zufällig sind, jedoch auf der Gesamtmenge der Anrufe einer konkreten Verteilung folgen.

Ein Anruf des Simulationsmodells hat bei seinem Eingang folgende Merkmale:

- **Zwischenankunftszeit** - Zeit zwischen Eingang des Anrufs und dem letztem vorher eingegangenen Anruf
- **potentielles Umsatzvolumen** - im Falle eines Vertriebs-Call-Centers ist dies die Geldmenge, die der Kunde umzusetzen wünscht

- **Routingrichtung** – bei der Routinggebietsstrategie hat der Anruf ein bestimmtes Call Center als Ziel

- **maximale Wartezeittoleranz** – müsste der Anrufer länger als seine Toleranz in einer Queue warten, so würde er auflegen

Im gewählten Simulationsmodell folgen die Ausprägungen der aufgezählten Merkmale einer merkmalsspezifischen Exponentialverteilung. Diese sollte hinsichtlich ihrer exponentiellen Rate parametrisiert sein, um Tests für unterschiedliche Verteilungen durchführen zu können. Durch die Rate wird die durchschnittliche Ausprägung des exponentialverteilten Merkmals vorgegeben. Geht beispielsweise im Mittel alle zwei Sekunden ein Anruf ein, so beträgt die Zwischenankunftsrate 0,5 (Anrufe pro Sekunde). Die Dichtefunktion der Zwischenankunftszeit hätte dann die Funktionsgleichung $f(x) = 0{,}5 * e^{-(0{,}5*t)}$.

Neben den aufgezählten Merkmalen, für die jeder Anruf von vornherein einen Wert zugewiesen bekommt, existieren weitere Anrufmerkmale, deren konkrete Ausprägungen erst beim Durchlaufen des Systems entstehen. Diese Merkmale sind:

- **Bearbeitungsdauer** – folgt einer standortabhängigen Exponentialverteilung, da jeder Standort seine spezifische mittlere Bearbeitungsrate besitzt

- **Deckungsbeitrag** – der Anteil vom erzielten Umsatz, den das Call Center erhält. Im Modell sollen lediglich die Umsätze betrachtet werde, weshalb der Deckungsbeitrag eines Anrufs gleich dem Umsatzvolumen ist. Bei Auflegern und Besetztfällen ist der Deckungsbeitrag immer Null. Der durchschnittliche Deckungsbeitrag pro Zeiteinheit ist dann die Summe der Umsätze während dieser Zeitdauer geteilt durch die Agentenzahl. Dies ist möglich, weil die Kosten im Modell ja nur vom Mitarbeitereinsatz abhängt, der im Simulationsmodell konstant gehalten werden soll.

- **Wartezeit** – Zeit, die der Anrufer ggf. in einer Queue verbracht hat, bevor er mit einem Agenten verbunden wurde

- **Wartefall** – falls der Anruf in einer Queue warten muss wahr, ansonsten falsch

- **Anrufabbruch** – falls der Anrufer auflegt wahr, ansonsten falsch

- **Besetztfall** – falls Anrufer geblockt wird wahr, ansonsten falsch

Neben diesen anrufspezifischen Merkmalen existiert für jede Bedienstation ein weiteres Merkmal, dass Auskunft über deren aktuellen Zustand gibt. Ein positiver Wert sagt

hierbei aus, dass sich gerade ein Anruf in Bedienung befindet, während ein negativer Wert für die Zeitdauer steht, die der Bedienplatz aktuell im unproduktiven Zustand auf einen neuen Anruf wartet.

Die eigentlich interessanten Leistungsmerkmale des Systems, werden durch Aggregation der Einzelmesswerte auf bestimmte Zeitintervalle ermittelt. Um beispielsweise die durchschnittliche Wartezeit eines Viertelstundenintervalls zu bestimmen, muss die Gesamtzeit, der im Zeitintervall anfallenden Wartezeiten der Anrufer ins Verhältnis zu den insgesamt ankommenden Anrufen desselben Zeitintervalls gesetzt werden.

Die Berechnung der Produktionsgrößen soll, wie in der Praxis größtenteils gebräuchlich, auf Basis von Viertelstundenintervallen erfolgen.

Die Größen, die im BSR-Modell sowohl für das Gesamtsystem, als auch für jedes einzelne Call Center berechnet und ausgewertet werden sollen, sind:

- **ankommende Anrufe** [ankAnr]
- **durchschnittliche Wartezeit** = Summe Wartezeiten / ankommende Anrufe
- **Wartewahrscheinlichkeit** = Summe Wartefälle / ankommende Anrufe
- **Centerauslastung** = (900 Sekunden * Bedienplätze - Summe Bereitzeiten) / (900 Sekunden * Bedienplätze)
- **Servicelevel(80/20)** = (abgenommene Anrufe im Servicelevel) / (ankommende Anrufe - Kurzzeitverzichter - Besetztfälle)
- **Summe Umsatz bzw. Deckungsbeitrag** [Db]

Da theoretisch unendlich viele Kombinationen kleinerer und größerer Call Center mit unterschiedlich großen Warteräumen möglich sind, muss mit dem Modell eine Auswahl getroffen werden, die im Hinblick auf die angestrebte Informationsgewinnung möglichst aussagefähig ist.

Um die Fragen aus Abschnitt 7.3 zumindest ansatzweise klären zu können, soll ein Modell mit drei Call Centern unterschiedlicher Größe als Anschauungsobjekt dienen. Das größte Call Center ist dabei mit 20 Bedienplätzen ausgestattet, das mittlere besitzt 14 und das kleinste 10 Bedienplätze. Da das große Call Center die doppelte Kapazität des kleinen hat, lassen sich bestimmte kapazitätsbedingte Effekte, über den Vergleich der beiden Call Center, möglicherweise besser einordnen. Das Modell soll sowohl mit

begrenzten Warteräumen, als auch für die Variante mit unbegrenzt großen Warteräumen simuliert werden. Der unbegrenzte Fall soll aber im Vordergrund stehen, da, wie in den vorangegangenen Kapiteln mehrmals erläutert, Besetztfälle aufgrund zu geringer Warteraumkapazität eher die Ausnahme darstellen und meist auf suboptimale Planung zurückgeführt werden können. Erreicht ein IB Call Center die angestrebten Servicewerte, so wie es mit der Planung angestrebt wird, so kommt es normalerweise auch nicht zu Besetztfällen.

Das Modell soll weiterhin sowohl mit als auch ohne Einsatz von BSR simuliert werden können, um den direkten Vergleich der beiden Varianten zu ermöglichen und den Effekt des BSR überhaupt einordnen bzw. abschätzen zu können. Zusätzlich muss außerdem noch ein einzelnes Call Center simuliert werden können, dass in seiner Anzahl der Bedienplätze der Summe der Bedienplätze der 3 BSR Call Center entspricht.

Die für Best Service Routing benötigte Simulation soll in MS Excel visuell dargestellt und die Ergebnisse in tabellarisch leicht verwertbarer Form gesammelt werden. Die Ablauflogik des Prozesses des Zufallsexperiments wird mit Hilfe von VBA implementiert werden.

Abbildung 8.1 zeigt die Visualisierung der Simulation.

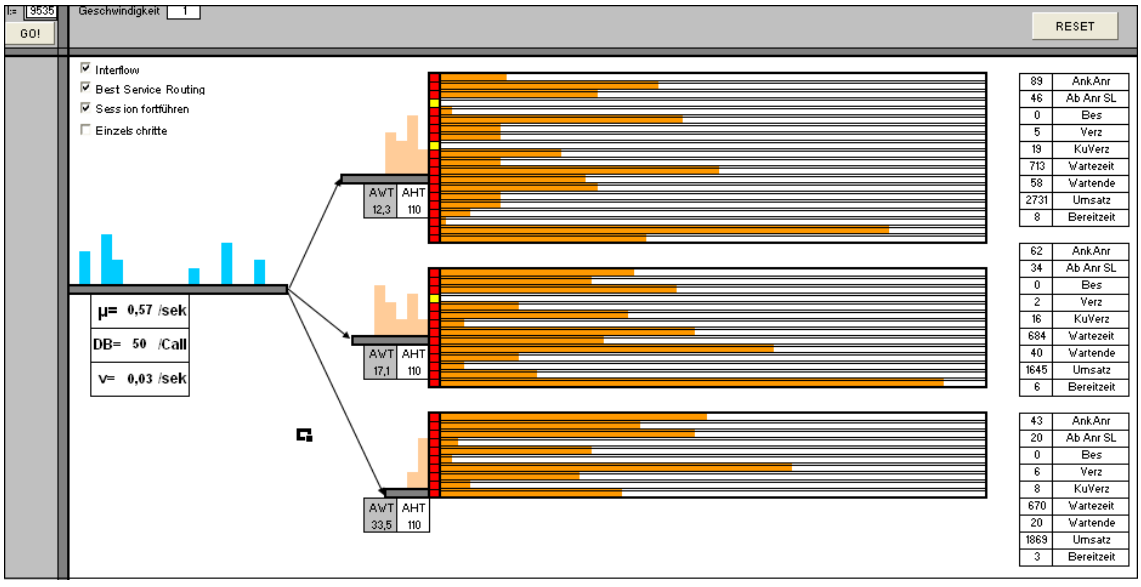


Abb. 8.1: Visualisierung der Simulation

Die Säulen im linken Bereich sind die ankommenden Anrufe, die am zentralen Verteilerpunkt auf die einzelnen Call Center verteilt werden. Wie lang ein Anruf dauern wird ist im Vorfeld noch nicht bestimmt. Die Höhe der Säulen gibt in diesem Fall nur

an, wie hoch der potentielle Umsatz sein wird, da ein Anrufer der die Nummer eines Versandhandels anwählt eine bestimmte Kaufabsicht mitbringt. Auch dieser potentielle Umsatz wird in diesem Modell mit einer Exponentialverteilung modelliert.

Die Pfeile im Bild, die vom Zentralverteilerpunkt abgehen, zeigen auf die drei Queues, die im Bild durch die waagerechten Balken am Ende der Pfeile dargestellt sind. In den Queues befinden sich wiederum die wartenden Anrufe, die wiederum durch diesmal eng aneinander liegende Säulen repräsentiert werden. Im Moment des Eintreffens in der Queue wird jedem Anruf eine maximale Wartezeittoleranz zugeordnet. Dauert die Anrufannahme länger als diese Toleranzzeit, so legt der Anrufer auf und verlässt automatisch die Queue. Die nachfolgenden wartenden Anrufe rücken dann im Anschluss nach.

Die drei großen rechteckigen Bereiche in der Mitte des Bildes stehen für die drei Call Center. Die unterschiedlich langen waagerechten Balken symbolisieren Anrufe, die sich gerade in Bedienung befinden. Je länger der Balken, umso länger dauert es, bis die zugehörige Bedieneinheit für einen neuen Anrufer frei wird. Ist ein Anruf abgearbeitet, so wird die Länge des Balkens Null und die Bedieneinheit wird frei. Sind in diesem Zeitpunkt gerade wartende Anrufer in der Queue, so gehen diese in die Bedienung über. Ansonsten wartet die nun freie Bedieneinheit, bis ein neuer Anruf zugeroutet wird. Die Zeitdauer, die die Bedieneinheit frei ist, wird ebenfalls gemessen und für jedes der drei Call Center kumuliert erfasst.

Das ganze System läuft in einem getackteten Zeitrhythmus. Nach einer festgesetzten Anzahl von Zeittakten erfolgt eine Abspeicherung der im abgelaufenen Intervall kumulierten Messwerte in eine Tabelle.

8.2 Die Inversionsmethode

Für die Simulation eines Warteschlangenmodells werden Zufallszahlen bestimmter Verteilungen benötigt, übliche Zufallszahlengeneratoren liefern aber nur stetige gleichverteilte Zufallszahlen im Intervall [0;1]. Soll eine bestimmte Verteilung simuliert werden, bedient man sich bei der Inversionsmethode deshalb eines Tricks. Man macht sich die Tatsache zunutze, dass der Wertebereich von Verteilungsfunktionen ebenfalls die reellen Zahlen im Intervall [0;1] umfasst. Zur Erinnerung: Der Funktionswert einer stetigen Verteilungsfunktion ist die Wahrscheinlichkeit, dass die Merkmalsausprägung eines beobachteten Sachverhalts kleiner oder gleich dem Abszissenwert ist.

Bei monoton steigenden Verteilungsfunktionen wie Exponentialverteilungen kann durch Bildung der Umkehrfunktion, ausgehend von den gleichverteilten Zufallszahlen, eine Zufallsvariable mit der geforderten Verteilung simuliert werden.

Abbildung 6.1 zeigt die Methode anhand der Funktionsgraphen. Die Funktion F^{-1} strebt in Richtung x=1 gegen unendlich.

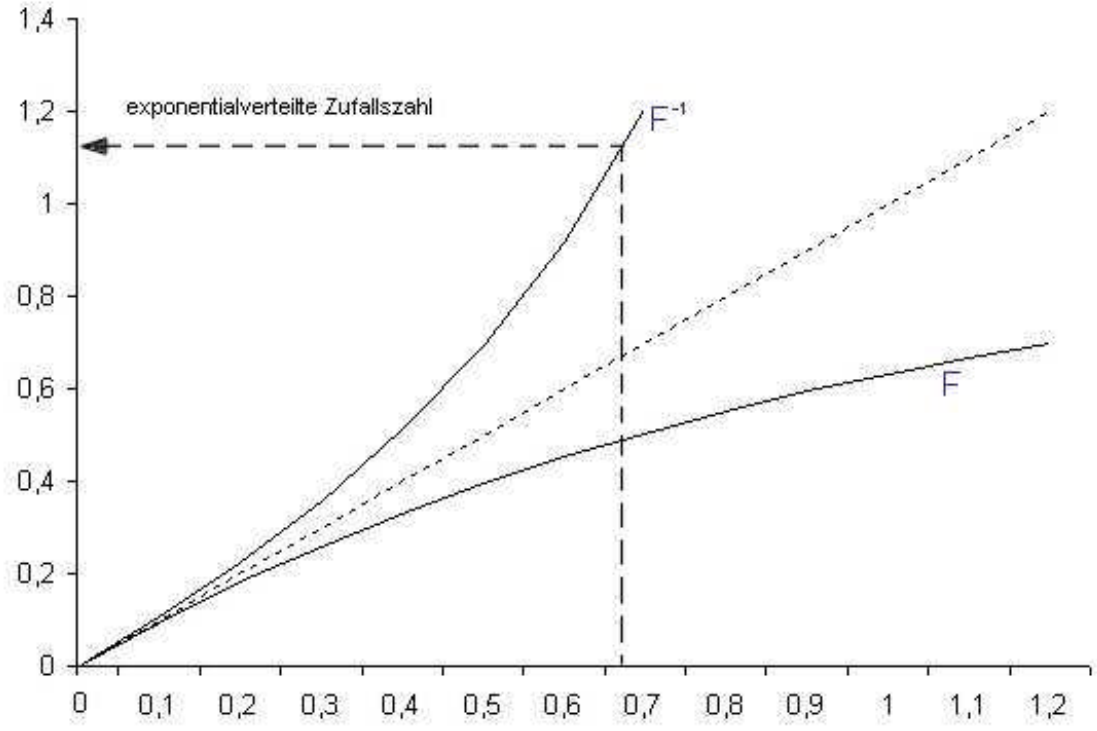


Abb. 8.2: Inversionsmethode für Exponentialfunktionen

Die Umkehrfunktion der Verteilungsfunktion $F(x) = 1 - e^{-(\lambda * x)}$ ist beispielsweise

$$x = F^{-1}(y) = -\frac{1}{\lambda}\ln(1-y)$$

Im Programm wird für y einfach die Zufallszahl zwischen Null und Eins des Zufallszahlengenerators eingesetzt. Der resultierende Funktionswert ist dann die exponentialverteilte Zufallsvariable.

Im hier verwendeten Simulationsmodell wurden die Zufallsvariablen der Zwischenankunftszeiten, der Anrufdauern, der Wartezeittoleranz und der potentiellen Umsätze mit Hilfe dieser Methode simuliert.

8.3 Der Prozessablauf der Simulation

Die Abbildung 8.3 zeigt den prinzipiellen Ablauf des Simulationsprozesses in einem EPK-Diagramm. Hieraus ist zu erkennen, dass ein System modelliert wurde, das eine Kombination von Routinggebiet, Overflow und BSR darstellt. Wie bereits erwähnt, greift BSR aber erst bei einem Anrufüberschuss, also wenn alle Bedienstationen belegt sind. Die Routinggebietsstrategie und das Überlaufsystem wurden deshalb vorgelagert, da auch für den Fall von zu wenigen Anrufen eine Routingstrategie erforderlich ist. Ab einem bestimmten Wert der Ankunftsrate werden die Routingvorgänge, die durch BSR bestimmt sind, aber zum Regelfall.

Jeder Anruf, der neu ins System kommt, hat trotzdem zuerst ein bestimmtes Call Center als Ziel. Die Verteilung für dieses Routing wurde diesmal aber nicht durch eine Exponentialfunktion modelliert, sondern es werden jedem Call Center fest vorgegebene Anteile des Gesamtanrufaufkommens zugeleitet, die dem Anteil der Bedienstationen des jeweiligen Centers von der Gesamtanzahl der Bedienstationen entsprechen. Auf diese Weise soll in etwa sichergestellt werden, dass bei Eintritt des BSR-Falles nicht einzelne Queues aufgrund ungleichmäßigen Routings permanent überlastet sind. Dies würde ansonsten eine gleichberechtigte Betrachtung der drei verschiedenen Call Center verhindern.

Um das Routing trotzdem möglichst zufällig zu gestallten wurde ein leistungsfähiger Mischalgorithmus verwendet, der die Reihenfolge, mit der die Routingzielvergabe erfolgt, ständig zufällig ändert.

Die EPK beschreibt nicht den kompletten Prozess, sondern soll hauptsächlich zum Verständnis der Ablauflogik dienen. Hinzuzufügen ist nur noch, dass beim Übergang eines Anrufs in die Bedienung auch noch einige Parameter wie etwa Umsatzvolumen und gewartete Zeit festgehalten und aufsummiert werden.

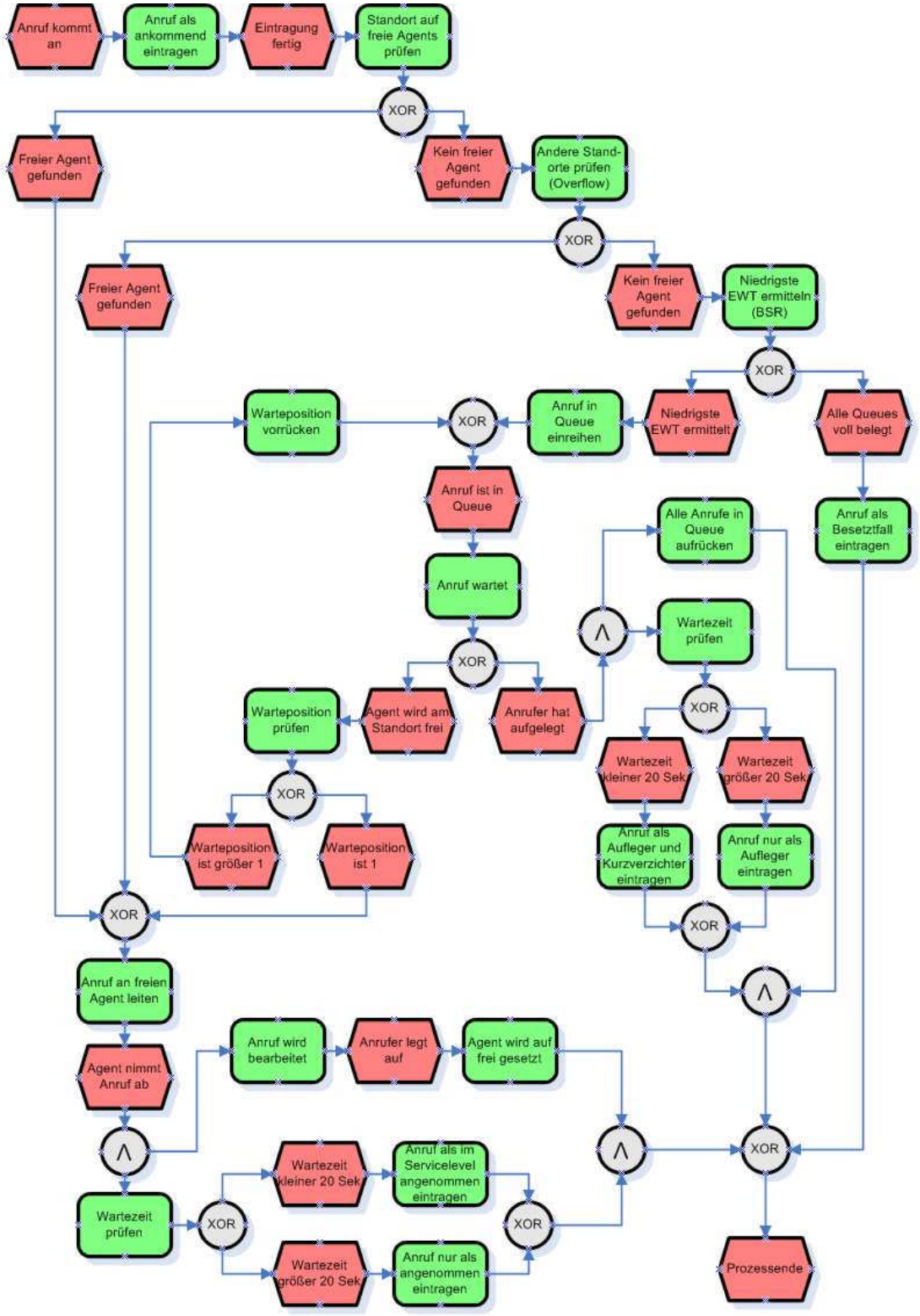


Abb. 8.3: Prozess eines eintreffenden Anrufs

8.4 Ergebnisse der Simulation

Im Ergebnis muss gesagt werden dass es mit BSR möglich ist, die Servicequalität im Gesamtsystem relativ deutlich zu erhöhen. Auch für jedes einzelne Call Center verbessern sich die Servicelevel durch BSR im Prinzip für jede simulierte Anrufankunftsrate. Weiterhin gehen die Wartezeiten zurück und die Auslastung der Bedieneinheiten wird erhöht.

Abbildung 8.4 zeigt die Veränderungen des gemessenen Servicelevels 80/20.

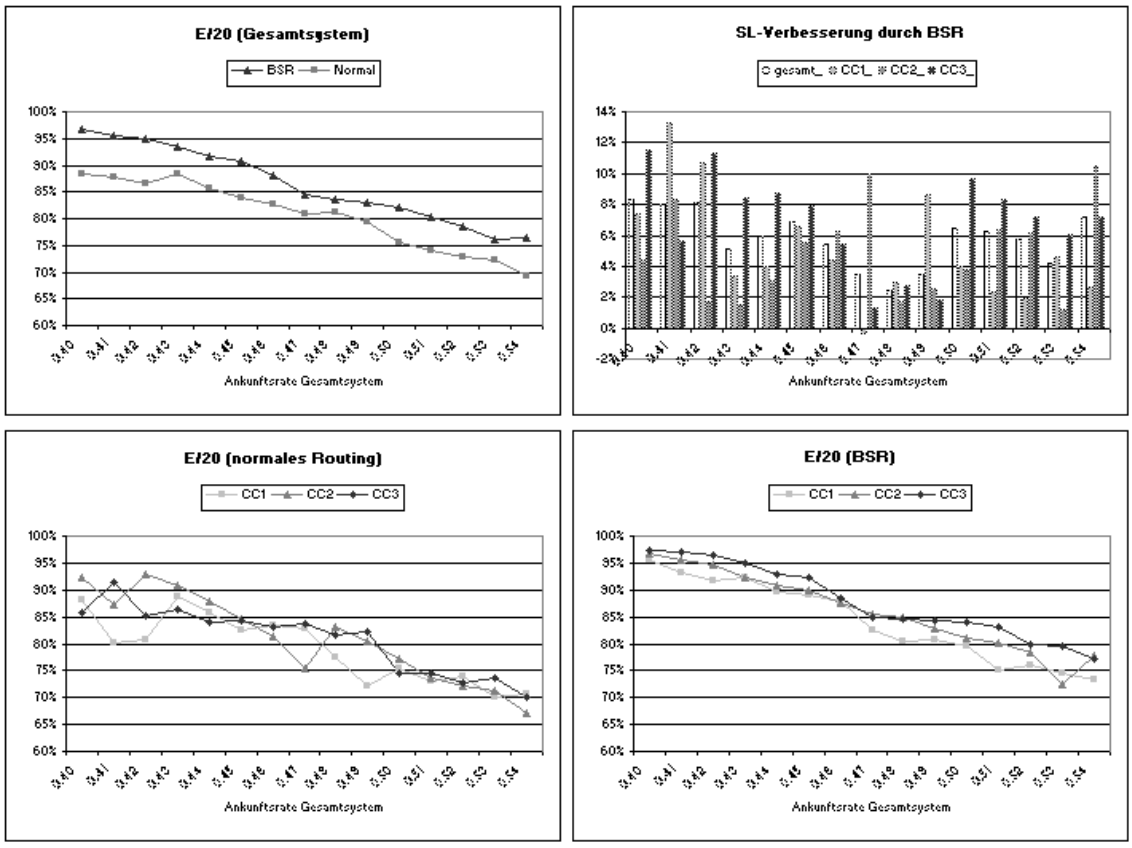


Abb. 8.4: Veränderungen des Servicelvels

Die positive Wirkung von BSR auf den Servicelevel ist deutlich zu sehen. Das vierte Diagramm verdeutlicht darüber hinaus die puffernde Wirkung des BSR bezüglich kurzzeitiger Schwankungen bestimmter Parameter wie beispielsweise den Bedienzeiten oder Zwischenankunftsraten. Während im linken unteren Diagramm relativ unruhige Zickzackkurven das Bild bestimmen, sind die Kurven im unteren rechten Diagramm relativ gleichmäßig und fast monoton. Allerdings kann trotz des relativ nahen Beisammenliegens der Servicelevelkurven im BSR-Fall schon ein leichter Vorteil der größeren Call Center vor den anderen abgelesen werden. Das würde letztendlich für die Praxis bedeuten, dass alle beteiligten Call Center versuchen müssten auf eine ähnliche Größe zu wachsen, damit alle möglichst denselben Servicelevel erreichen. Die größeren

Call Center würden ansonsten immer nur versuchen, den Servicelevel durch ihren Mitarbeitereinsatz genau zu treffen, was für die kleineren Call Center bedeuten würde, dass sie ihn ohne Größenwachstum nie erreichen

Auch bei den Umsatzkurven ist die Pufferwirkung des BSR deutlich zu sehen. Auch hier liegen die Kurven enger zusammen als im Fall ohne BSR.

Interessant zu sehen ist, dass im BSR-Fall bei einer Ankunftsrate von ungefähr λ=0,55 für das Gesamtsystem und auch alle einzelnen Call Center, ein Optimum erreicht zu sein scheint. Die Kurven gehen nach λ=0,55 größtenteils in einen Abwärtstrend über. Im Fall ohne BSR ist dieses Optimum schon etwas früher, bei ungefähr λ=0,53, erreicht.

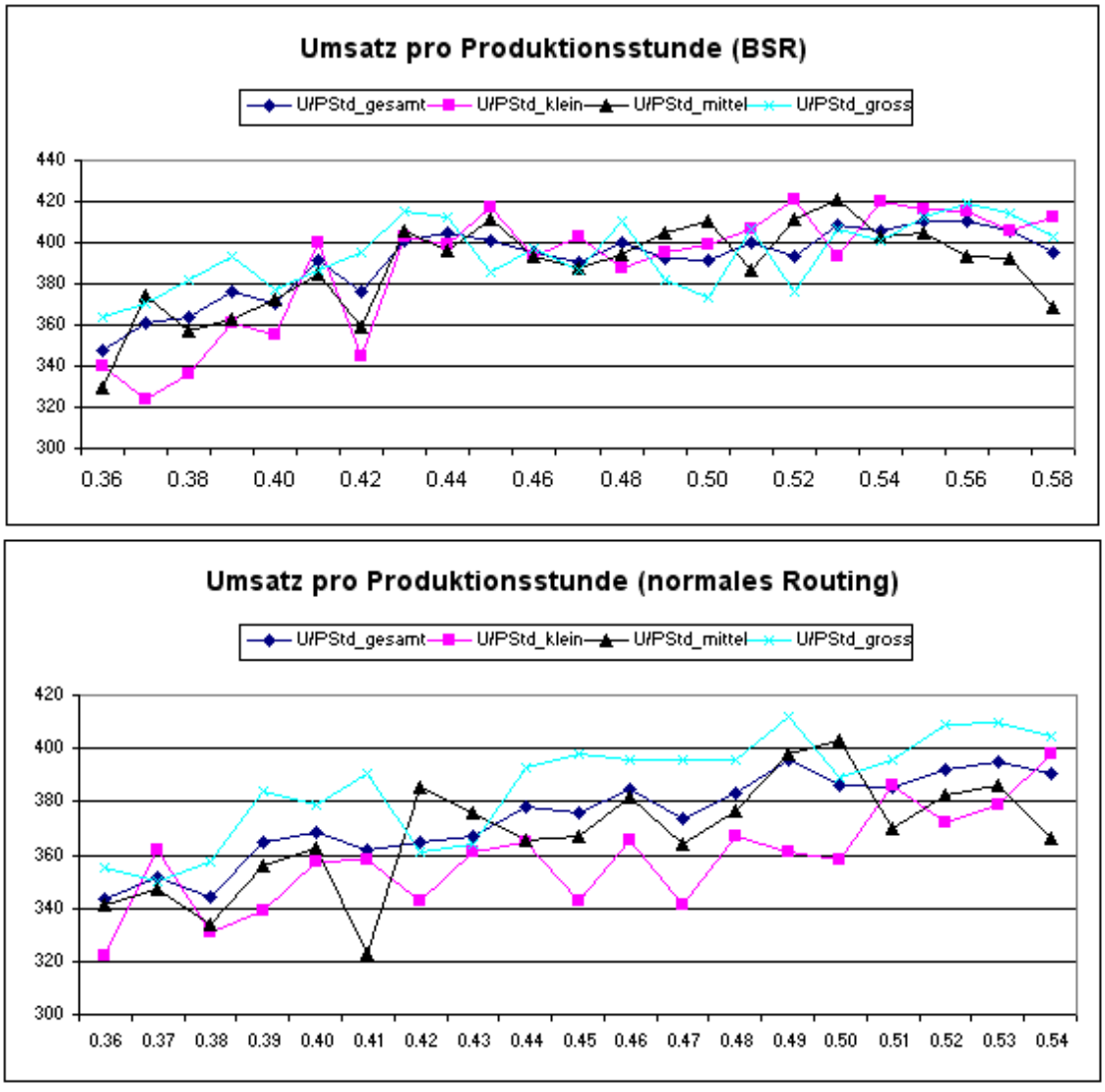


Abb. 8.5: Umsatzkurven

Die wesentlichste Aussage ist nun aber, dass wir als Grundlage für das finale Optimierungsmodell sagen können, dass die Anrufverteilung fast vollständig proportional zum jeweiligen Größenanteil des Call Centers, gemessen am Gesamtsystem ist. Ein Call-Center mit doppelt so vielen Bedienplätzen bekommt also auch doppelt so viele Anrufe zugeroutet, wie ein anderer Teilnehmer mit nur halber Besetzung. In der Testreihe unter BSR-Bedingungen erhielt das große Call Center durchschnittlich 45,3% der Anrufe bei einem Größenanteil von 45,4% bezogen auf die Gesamtanzahl der Bedienplätze. Das mittlere Cal Center erhielt 31,6% der Anrufe bei einem Größenanteil von 31,8% und das kleine Call Center mit 10 Bedienplätzen erhielt 23% der ankommenden Anrufe bei einem Größenanteil von 22,8%. Die vollständigen Datenreihen mit allen gemessenen Werten bis zu einer Ankunftsrate von λ=0,54 zeigt die Tabelle 8.1.

Tab. 8.1: Ergebnisreihen der BSR Simulation

		λ																		
Messwert	Routing	0.36	0.37	0.38	0.39	0.40	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.50	0.51	0.52	0.53	0.54
E20_gesamt	BSR	99%	98%	97%	96%	97%	96%	95%	93%	92%	91%	88%	85%	84%	83%	82%	80%	78%	76%	76%
E20_gesamt	Normal	95%	93%	93%	93%	88%	88%	87%	88%	86%	84%	83%	81%	81%	79%	76%	74%	73%	72%	69%
E20_klein	BSR	99%	97%	95%	94%	96%	93%	92%	92%	90%	89%	88%	82%	80%	81%	80%	75%	76%	75%	73%
E20_klein	Normal	92%	92%	93%	89%	88%	80%	81%	89%	86%	83%	83%	83%	77%	72%	76%	73%	74%	70%	71%
E20_mittel	BSR	98%	98%	98%	96%	97%	95%	95%	92%	91%	90%	87%	85%	85%	83%	81%	80%	78%	73%	78%
E20_mittel	Normal	96%	95%	93%	92%	92%	87%	93%	91%	88%	85%	81%	76%	83%	80%	77%	74%	72%	71%	67%
E20_gross	BSR	99%	99%	98%	97%	97%	97%	96%	95%	93%	92%	88%	85%	85%	84%	84%	83%	80%	80%	77%
E20_gross	Normal	96%	92%	94%	95%	86%	91%	85%	86%	84%	84%	83%	84%	82%	82%	74%	75%	73%	74%	70%
U/PStd_gesamt	BSR	348	361	364	376	371	392	376	401	404	401	395	391	400	392	391	400	393	408	406
U/PStd_gesamt	Normal	343	352	344	365	369	362	365	367	378	376	384	373	383	396	386	385	392	395	391
U/PStd_klein	BSR	340	324	336	361	356	400	344	403	399	417	393	403	387	395	399	407	421	394	420
U/PStd_klein	Normal	322	362	331	339	358	358	343	361	365	343	365	341	367	361	358	386	372	379	398
U/PStd_mittel	BSR	330	375	357	363	372	385	359	406	396	411	394	388	395	405	411	386	411	421	403
U/PStd_mittel	Normal	342	347	334	356	362	323	386	376	366	367	382	364	376	398	403	370	382	386	366
U/PStd_gross	BSR	364	370	382	394	377	386	395	416	413	386	397	387	411	382	373	407	376	406	401
U/PStd_gross	Normal	355	350	358	384	378	391	361	364	393	398	396	396	395	412	389	396	409	410	404
Auslast_gesamt	BSR	0,88	0,9	0,93	0,94	0,94	0,95	0,94	0,97	0,97	0,97	0,98	0,98	0,99	0,99	0,99	0,99	1,00	0,99	0,99
Auslast_gesamt	Normal	0,84	0,83	0,85	0,87	0,88	0,91	0,9	0,91	0,93	0,94	0,93	0,93	0,94	0,95	0,95	0,96	0,96	0,96	0,97
Auslast_klein	BSR	0,87	0,87	0,92	0,94	0,93	0,94	0,95	0,96	0,97	0,97	0,98	0,98	0,98	0,99	0,99	0,99	1,00	0,99	0,99
Auslast_klein	Normal	0,81	0,77	0,81	0,83	0,86	0,89	0,89	0,89	0,9	0,91	0,9	0,89	0,92	0,93	0,91	0,94	0,94	0,95	0,94
Auslast_mittel	BSR	0,86	0,9	0,92	0,94	0,94	0,95	0,94	0,96	0,97	0,97	0,98	0,98	0,99	0,99	0,99	0,99	1,00	0,99	0,99
Auslast_mittel	Normal	0,82	0,8	0,83	0,86	0,85	0,9	0,88	0,88	0,92	0,94	0,94	0,94	0,92	0,94	0,95	0,96	0,96	0,95	0,96
Auslast_gross	BSR	0,89	0,9	0,93	0,95	0,95	0,96	0,95	0,97	0,98	0,97	0,98	0,99	0,99	0,99	0,99	0,99	1,00	0,99	0,99
Auslast_gross	Normal	0,88	0,87	0,89	0,89	0,91	0,92	0,92	0,95	0,96	0,95	0,95	0,95	0,97	0,97	0,97	0,98	0,97	0,97	0,98
AWT_gesamt	BSR	1,23	1,84	2,84	3,23	3,31	3,63	3,97	5,37	6,01	6,06	7,62	8,46	8,88	9,34	9,49	10,5	11	10,8	11,2
AWT_gesamt	Normal	3,73	4,34	4,17	4,35	7,21	6,22	7,65	8,21	8,38	8,46	9,12	8,49	9,37	10,1	11,1	11,7	11,3	11,4	12,7
AWT_klein	BSR	1,89	2,8	4,21	4,75	5,01	5,06	6,08	7,96	9,03	8,59	10,7	12,1	12,6	13,4	13,1	14,7	16,1	15,2	15,5
AWT_klein	Normal	5,24	5,78	5,91	6,21	10,7	9,33	11,6	12,6	12,3	12,4	13,6	12,7	14,1	15,2	16,6	17,8	17	17,6	18,5
AWT_mittel	BSR	1,29	1,9	2,98	3,53	3,5	3,77	4,23	5,65	6,31	6,51	8,03	8,7	9,29	9,69	10,2	11	11,6	11,5	11,8

AWT_mittel	Normal	4,1	4,46	4,57	4,65	7,26	6,39	8,07	8,41	9,04	9,07	9,7	9,03	10	10,9	11,9	12,5	12,2	11,8	13,6
AWT_gross	BSR	0,88	1,34	2,07	2,3	2,37	2,76	2,82	3,92	4,36	4,47	5,69	6,36	6,64	6,98	7,11	8	8,08	7,94	8,47
AWT_gross	Normal	2,78	3,58	3,1	3,27	5,53	4,66	5,52	6	6,14	6,19	6,64	6,2	6,72	7,21	7,94	8,26	8,13	8,22	9,26
Anrufe_gesamt	BSR	3201	3368	3480	3542	3505	3731	3714	3892	4029	4119	4158	4333	4304	4430	4488	4570	4739	4753	4836
Anrufe_gesamt	Normal	3289	3321	3398	3547	3604	3722	3751	3845	4030	4060	4238	4247	4345	4480	4512	4644	4719	4821	4871
AnrAnteil_klein	BSR	22%	21%	23%	22%	22%	24%	22%	22%	22%	23%	24%	23%	24%	23%	24%	24%	23%	24%	24%
AnrAnteil_klein	Normal	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%
AnrAnteil_mittel	BSR	31%	33%	31%	31%	32%	32%	31%	32%	32%	31%	32%	32%	32%	32%	31%	32%	32%	31%	32%
AnrAnteil_mittel	Normal	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%	31%
AnrAnteil_gross	BSR	47%	46%	46%	47%	46%	44%	47%	46%	46%	46%	45%	44%	44%	45%	45%	44%	45%	45%	44%
AnrAnteil_gross	Normal	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%
Δ_E20_klein		0,04	0,05	0,04	0,03	0,08	0,08	0,08	0,05	0,06	0,07	0,05	0,04	0,02	0,03	0,06	0,06	0,06	0,04	0,07
Δ_E20_gesamt		0,07	0,05	0,03	0,05	0,07	0,13	0,11	0,03	0,04	0,07	0,04	-0	0,03	0,09	0,04	0,02	0,02	0,05	0,03
Δ_E20_mittel		0,02	0,03	0,05	0,04	0,04	0,08	0,02	0,02	0,03	0,06	0,06	0,1	0,02	0,02	0,04	0,06	0,06	0,01	0,1
Δ_E20_gross		0,04	0,07	0,04	0,02	0,12	0,06	0,11	0,08	0,09	0,08	0,05	0,01	0,03	0,02	0,1	0,08	0,07	0,06	0,07

Natürlich sind die bis hierhin getätigten Analysen und die daraus abgeleiteten Aussagen nicht unbedingt als allgemeingültig anzusehen. Bei der unendlichen Vielzahl von denkbaren Systemkonfigurationen ist dies mit Hilfe einer Simulation auch niemals vollständig möglich. Trotzdem sind einige Ergebnisse recht überzeugend und zumindest wenn im betrachteten System vorausgesetzt wird, dass alle Teilnehmer ungefähr dieselben durchschnittlichen Bearbeitungszeiten aufweisen, kann von einem ähnlichen realen Verhalten ausgegangen werden, wie dies in diesem Modell gezeigt wurde.

9 Das finale Optimierungsmodell des BSR

Durch die Ergebnisse der Simulation kann nun das endgültige Modell für die Optimierung des Mitarbeitereinsatzes eines einzelnen Call Centers im BSR-Modell vorgenommen werden.

In den jeweiligen Formeln für das Vertriebs- und für das Dienstleistungs-Call-Center (Formel 6.3 und 6.4) muss der Parameter der ankommenden Anrufe überall gegen eine zusätzliche Formelerweiterung ausgetauscht werden.

Die Menge der ankommenden Anrufe ist nun nicht mehr nur von der eigenen Anzahl eingesetzter Mitarbeiter abhängig, sondern zusätzlich auch von der geschätzten Mitarbeiteranzahl der anderen Mitbewerber. Da wir nun voraussetzen, dass das Verhältnis aus den am Standort erhaltenen Anrufen zum Gesamtanrufaufkommen im System gleich dem Verhältnis aus den am Standort eingesetzten Mitarbeitern zu den insgesamt eingesetzten Mitarbeitern ist, können wir folgendes konstruieren:

$$AnkAnr(s) = AnkAnr(ges) * \frac{c}{c + c(andere)} \quad (9.1)$$

Die Gesamtanzahl der Mitarbeiter der anderen Standorte c(andere) berechnet sich aus der Differenz aller Mitarbeiter c(ges) im System und den eigenen eingesetzten Mitarbeitern am Standort. Vor der Optimierung muss c(ges) anhand des gemessenen Verhältnisses der Anrufverteilung und des eigenen gemessenen Mitarbeitereinsatzes c(s) bestimmt werden.

$$c(ges) = \frac{AnkAnr(ges) * c(s)}{AnkAnr(s)}$$

Es gilt:

$$c(andere) = c(ges) - c(s)$$

In den Optimierungsformeln aus Kapitel 6 wird nun also einfach der Parameter AnkAnr gegen AnkAnr(s) aus Formel 9.1 ausgetauscht.

Bei der Optimierung lässt man c(andere) in Formel 9.1 dann konstant, da ja nur bestimmen werden soll, wie sich die Gewinne bzw. der Servicelevel verändern, wenn die Mitarbeiterzahl am eigenen Standort variiert. So soll in etwa ermittelt werden, wie man in einem Unterdeckungsfall allein die überschüssigen Mengen gewinnmaximal ausgenutzt hätte.

Im umgekehrten Fall ist das vom Prinzip her das gleiche. Man optimiert um herauszubekommen, wie viele Mitarbeiter weniger gebraucht worden wären, um den optimalen Servicelevel oder den maximalen Gewinn zu erreichen.

Die Optimierung an sich läuft dann wie schon vorab beschrieben durch schrittweises Erhöhen der Agentenanzahl c.

Die so erhaltenen optimale Zeitreihen der Mitarbeiteranzahl liefern dann die Grundlage für eine Prognose derselben. Hier würde sich das Verfahren der exponentiellen Glättung zweiter Ordnung anbieten, um auch gegebenenfalls einen Trend berücksichtigen zu können. Denkbar wäre aber auch die Bildung eines gewichteten Mittelwerts, bei dem beispielsweise umgekehrt proportional zum Abstand des Optimalwertes zum realen Messwert gewichtet wird. Hier bleibt eine Menge Spielraum für Weiterentwicklungen, denn das Modell ist so dynamisch und komplex, dass eine wirklich annähernd optimale Lösung noch in weiter Ferne scheint.

10 Ausblick

Durch noch genauere Analyse der Wechselbeziehungen im BSR und durch Einbeziehung von Aussagen zur Messwertstreuung sowie der Prognosequalität, kann das System an vielen Stellen erweitert werden.

Die Einbeziehung von KNN könnte darüber hinaus eine Möglichkeit sein um ein Lernverhalten im System zu implementieren, über welches beispielsweise die erwartete Reaktion der Mitbewerber auf signifikante Unter- oder Überbesetzungen der vergangenen Perioden einbezogen werden könnte. Beispielsweise könnten bestimmte Zusatzparameter in der Zielfunktion implementiert werden, die variiert werden, um die Prognosegüte weiter zu erhöhen.

Das hier vorgestellte Verfahren ist somit nur als erster Schritt zu sehen, der ein Mittel liefern soll, überhaupt Voraussagen für ein so dynamisches System wie das BSR zu tätigen.

Anhang

Datenreihe eines realen Call Centers ohne BSR

Zeit	E20	AnkAnr Vdnxx02	Bes	Verz	KuVerz	AbAnr	AbAnrSlv	mAwz	mKz	mGz	Pl1/4h
Gesamt:	85,40	56671	77	4780	2753	104741	91258	9,16	1,48	48,6	7745
09:45	88,73	1503	0	53	30	1440	1307	7,5	1,3	47,8	105,9
10:30	81,13	1650	0	62	39	1548	1307	11,6	1,4	47,3	114,9
10:45	82,68	1657	0	79	40	1564	1337	12,5	1,5	50	116,5
11:00	85,63	1568	0	63	37	1544	1311	11,7	1,4	49,1	115,1
11:30	86,71	1600	0	76	42	1509	1351	10	1,4	46,9	107,6
12:30	80,81	1082	0	43	24	1026	855	10,7	1,4	50,9	78,84
13:00	83,5	1040	0	54	28	983	845	13,6	1,5	48,7	67,97
13:15	86,69	1084	0	66	40	1020	905	13	1,7	49	68,85
15:30	89,95	1073	0	39	18	1028	949	9,3	1,5	48,4	76,7
21:45	87,76	157	0	13	10	140	129	6	1,5	46,3	11,96
07:15	80	31	0	2	1	28	24	8,7	1,5	47,4	4
12:30	87,88	343	0	19	13	313	290	8,2	1,6	48,1	24,27
14:45	83,85	303	1	17	12	285	244	9,6	1,6	50,9	20,66
15:00	83,69	296	0	23	14	275	236	7,9	1,5	51,3	20,59
15:30	84,03	328	0	21	15	296	263	6,7	1,4	49,3	23,53
15:45	89,31	294	0	5	4	277	259	4,8	1,3	50,9	24
16:00	83,75	331	0	18	11	308	268	8,9	1,4	54,4	24,07
17:00	86,49	354	0	34	21	336	288	8,8	1,2	53,6	26,98
17:30	89,17	369	0	27	18	370	313	8,7	1,2	52,4	27,92
07:15	80	31	0	2	1	28	24	8,7	1,5	47,4	4
12:30	87,88	343	0	19	13	313	290	8,2	1,6	48,1	24,27
14:45	83,85	303	1	17	12	285	244	9,6	1,6	50,9	20,66
15:00	83,69	296	0	23	14	275	236	7,9	1,5	51,3	20,59
15:30	84,03	328	0	21	15	296	263	6,7	1,4	49,3	23,53
15:45	89,31	294	0	5	4	277	259	4,8	1,3	50,9	24
16:00	83,75	331	0	18	11	308	268	8,9	1,4	54,4	24,07
17:00	86,49	354	0	34	21	336	288	8,8	1,2	53,6	26,98
17:30	89,17	369	0	27	18	370	313	8,7	1,2	52,4	27,92
08:00	87,41	1035	0	42	26	991	882	9,2	1,5	44,9	65,28
11:30	82,89	1847	0	74	35	1759	1502	11,4	1,5	49,6	130,1
12:30	89,04	1253	0	37	21	1207	1097	7,1	1,4	48,7	90,81
14:30	86,83	1441	0	46	29	1364	1226	10	1,6	49,3	107,4
15:30	86,21	1504	0	58	32	1430	1269	9,6	1,4	49,5	104,9
20:00	84,52	433	0	21	13	430	355	9,1	1,4	51,4	32,25
21:30	88,64	186	0	13	10	168	156	6,1	1,5	55,4	15,17
22:30	88,1	89	0	7	5	81	74	5,4	1,3	45	8
09:15	82,98	1634	0	66	36	1539	1326	11,4	1,5	50	109,2
10:30	88,29	1727	0	73	44	1641	1486	10,5	1,5	47,8	126,1
11:15	87,3	1622	0	35	23	1544	1396	7,6	1,4	50,2	116
12:00	83,33	1297	0	41	19	1232	1065	11,1	1,4	48,7	90,33
17:45	82,71	943	0	70	35	875	751	11,2	1,5	48,7	61,53
22:00	87,3	132	0	9	6	122	110	8,7	1,5	53,5	10,4
22:15	82,79	125	3	5	3	119	101	9,7	1,1	49,6	10
23:00	82,35	19	0	12	2	15	14	3,1	1,2	40,9	1,45
08:00	82,85	776	0	27	12	741	633	8,1	1,4	45,3	51,73

12:00	82,47	1298	0	39	20	1227	1054	11,2	1,4	51,3	94,88
14:45	89,85	1344	0	59	34	1280	1177	11	1,5	48,1	98,57

Literaturverzeichnis

Adeli, H.; Hung, S. (1995): Machine Learning – Neural Networks, Genetic Algorithms and Fuzzy Systems. New York – Chichester et al.

Borgwardt, K.-H. (2001): Optimierung Operations Research Spieltheorie – Mathematische Grundlagen. Basel – Boston - Berlin

Bossert, M.; Breitbach, M. (1999): Digitale Netze. Stuttgart - Leipzig

Burda, A.; Färber, G. (1995): Das große Buch zu Delphi. Wien

Chen, P. (1976): The Entity-Relationship-Model – Towards a Unified View of Data. ACM Transactions on Database Systems, Band 1, Nr. 1, S. 9-36

Dumke, R (2001): Software Engineering. 3. Aufl., Wiesbaden

Fujimoto, R. M. (1999): Parallel and Distributed Simulation Systems. Wiley-Interscience. New York

Helbert, S.; Stolletz, R. (2003): Call Center Management in der Praxis: Strukturen und Prozesse betriebswirtschaftlich optimieren. Berlin – Heidelberg

Heuer, A.; Saake, G; Sattler, K.-U. (2001): Datenbanken-kompakt. Paderborn

Jobson, J.D. (2005): Applied Multivariate Data Analysis. New York – Berlin - Heidelberg

King, W.; Hufnagel, E.; Grover, V. (1989): Using Information Technology for Competitive Advantage. Oxford – New York et al.

Patterson, D. (1997): Künstliche neuronale Netze. 2. Aufl., München u. a.

Rautenstrauch, C. (2007): Grundlagen der Wirtschaftsinformatik, o. Jg, Vorlesungsskript, S. 55

Reinhart, M. (1995): Relationales Datenbankdesign. München

Sniedovich, M. (1997): Dynamic Programming. New York – Basel – Hong Kong

Stier, W. (2001): Methoden der Zeitreihenanalyse. Berlin et al.

Internet-Adressen

TSE Hamburg (2007): Call Center Glossar http://www.tse-hamburg.de/papiere/call%20center%20und%20telekommunikation/ACDGlossar.html. 28. Juni 2007

Pandis, I. (2006): Database Applications http://www.cs.cmu.edu/~ipandis/15415/F06/HW/hw10.htm. 01. Juli 2007

Wikipedia (2007), Agner Krarup Erlang, http://de.wikipedia.org/wiki/Agner_Krarup_Erlang

Abschließende Erklärung

Ich versichere hiermit, daß ich die vorliegende Diplomarbeit selbständig, ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Magdeburg, den 02. September 2008