3. Personalbedarfsermittlung als Basis der Personalbestands- und Personaleinsatzplanung

Call Center sind mit einer im Zeitablauf stark variierenden Nachfrage konfrontiert. Zudem können sie weder den Zeitpunkt und das Volumen der Nachfrage substanziell beeinflussen, noch haben sie die Möglichkeit, auf Lager zu produzieren. Demnach müssen ausreichend Agenten zur Verfügung gestellt werden, um die Kundenanfragen zum Zeitpunkt ihres Auftretens mit der gewünschten Maßzahl der Kundenzufriedenheit beantworten zu können. Zu diesem Zweck ist der Personalbedarf zu ermitteln, der die Anzahl der benötigten Agenten für einen bestimmten Zeitraum bei einer antizipierten Nachfrage bestimmt. Ausgehend von dem Personalbedarf wird der Personalbestand bestimmt, d.h. die Anzahl der vertraglich gebundenen Agenten mit deren unterschiedlichen Arbeitszeiten, um den Bedarf zu befriedigen. Die Personalbestandsbestimmung kann jedoch nicht losgelöst von der späteren Nutzung des Personals im Rahmen der zeitlichen Verteilung der Arbeitszeit, d.h. der Personaleinsatzplanung erfolgen. Der Personalbedarf bildet die Voraussetzungen für die Personalbestands- und Personaleinsatzplanung und bestimmt somit auch maßgeblich die Qualität der Planung.

Das dritte Kapitel befasst sich mit den Voraussetzungen für die Personalbestands- und aggregierte Personaleinsatzplanung. Diese beruhen auf der Bestimmung des Personalbedarfs. Im Abschnitt 3.1 richtet sich das Augenmerk zunächst auf die Determinanten des Personalbedarfs. Eine Determinante, die bei Call Centern den Personalbedarf maßgeblich bestimmt, besteht – wie bereits mehrfach beschrieben – in der Nachfrage nach den Dienstleistungen. Verschiedene Möglichkeiten, die Nachfrage im Call Center zu prognostizieren, werden im Abschnitt 3.2 vorgestellt. Vor der Beschreibung der Methoden der Personalbedarfsermittlung findet im Abschnitt 3.3 eine Erörterung der Anforderungen an diese Methoden statt, um eine geeignete Methode für kleine Call Center mit homogenen Agenten und Kunden auszuwählen. Im Anschluss daran erfolgt in den Abschnitten 3.4 und 3.5 die Darstellung der Methoden zur Ermittlung des kurz- und mittelfristigen Personalbedarfs. Bei der kurzfristigen Personalbedarfsermittlung wird zunächst im Abschnitt 3.4.1 die Bestimmung des Personalbedarfs mittels der Regressionsanalyse untersucht und diese hinsichtlich ihrer Anwendungsmöglichkeiten in Call Centern geprüft. Der sich anschließende Abschnitt 3.4.2 widmet sich den Warteschlangenmodellen, die bei Call Centern normalerweise zur Personalbedarfsermittlung im operativen Bereich eingesetzt werden. Abschnitt 3.5 beschäftigt sich mit der mittelfristigen Personalbedarfsermittlung. Dabei ist

u.a. zu eruieren, inwieweit Warteschlangenmodelle für die mittelfristige Personalbedarfsermittlung einsetzbar sind. Zu diesem Zweck werden die Möglichkeiten genutzt, den Personalbestand anhand des periodenweise bestimmten und aggregierten Personalbedarfs zu berechnen und in einen mittelfristigen Personalbedarf zu transformieren.

3.1 Determinanten des Personalbedarfs

Bei der Ermittlung des Personalbedarfs geht es grundsätzlich darum, zuverlässige Determinanten bzw. Bezugsgrößen zu identifizieren, die den Personalbedarf maßgeblich bestimmen. Für Call Center sind die folgenden Determinanten ausschlaggebend für den Personalbedarf:[121]

- das Anrufvolumen,
- die Bearbeitungszeit eines Anrufes sowie
- die Arbeitszeit eines Agenten.

Die Nachfrage, die sich im *Anrufvolumen* konkretisiert, muss zum Zeitpunkt ihres Auftretens befriedigt werden. Dies geschieht, indem kontinuierlich ausreichend Agenten zur Verfügung gestellt werden. Dies wurde bereits im Abschnitt 2.4 ausführlich beschrieben. Die von außen eingehende Nachfrage in jeder Periode stellt somit einen wesentlichen Bestimmungsfaktor des Personalbedarfs dar. Das Anrufvolumen für Zeitintervalle unterschiedlicher Länge kann mit Hilfe von verschiedenen Prognoseverfahren vorhergesagt werden. Diese sind Gegenstand des Abschnitts 3.2.

Die *Bearbeitungszeit eines Anrufes* setzt sich aus der Gesprächszeit und der Nachbearbeitungszeit zusammen. Die Länge der Bearbeitungszeit spielt bei der Bestimmung des Personalbedarfs eine bedeutende Rolle. Lange Bearbeitungszeiten führen zu einer höheren Inanspruchnahme des Faktors Personal als kurze Bearbeitungszeiten und haben bei identischem Anrufvolumen eine größere Anzahl an Agenten zur Folge. Die Bearbeitungszeit kann in Abhängigkeit von der Tageszeit variieren. Die Vergangenheitsdaten bezüglich der durchschnittlichen Gesprächszeit sind Bestandteil der Aufzeichnungen der ACD-Anlage und liegen folglich vor. Die durchschnittliche Nachbearbeitungszeit ist nicht immer in der ACD-Anlage ablesbar, sie kann aber anhand von Stichproben geschätzt werden. Sofern das Leistungsprogramm des Call Centers keinen Änderungen unterliegt, können die Daten der

[121] Ohne spezifischen Branchen- oder Unternehmensbezug werden das Leistungsprogramm, die Arbeitsproduktivität und die Arbeitszeit als primäre Determinanten zur Bestimmung des Personalbedarfs genannt. Vgl. Mag (1998), S. 70-72 und Kossbiel (1992), Sp. 1598 f.

ACD-Anlage als Grundlage verwendet werden. Andernfalls ist eine Anpassung notwendig.

Die *Arbeitszeit eines Agenten* gehört nur bei der mittelfristigen Bestimmung des Personalbedarfs zu den Determinanten. Bei der Bestimmung der Anzahl der zu beschäftigenden Agenten ist die Arbeitszeit eine Determinante. Sie gibt Auskunft über die Anzahl der Arbeitsstunden, die der Agent dem Call Center in dem betrachteten Zeitraum zur Verfügung steht und somit Kundenanfragen befriedigen kann. Bei der kurzfristigen Personalbedarfsbestimmung hingegen entfällt sie als Bezugsgröße. Der kurzfristige Personalbedarf im Call Center beschreibt lediglich die Anzahl an Personen, die in einem kurzen Intervall von 15 bis 60 Minuten zur Aufgabenerfüllung benötigt werden. Diese Anzahl an Agenten ist jedoch unabhängig von deren Arbeitszeit, da das Zeitintervall des kurzfristigen Personalbedarfs einen Bruchteil der täglichen Arbeitszeit eines Agenten ausmacht.

Die Arbeitszeit kann zwischen Arbeitgeber und Agenten unter Berücksichtigung der gesetzlichen, tariflichen und betrieblichen Rahmenbedingungen vereinbart werden. Die Rahmenbedingungen wurden bereits im Abschnitt 2.5.3.3 beschrieben. Die Arbeitszeit eines Agenten lässt sich in die nominelle, die reale und die produktiv nutzbare Arbeitszeit unterteilen. Abbildung 3.1 gibt einen Überblick über deren Zusammenhang. Die *nominelle* Arbeitszeit umfasst die vertraglich vereinbarte Arbeitszeit in einem Zeitraum. Die *reale* Arbeitszeit eines anwesenden Agenten entspricht seiner nominellen Arbeitszeit erhöht um seine Mehrarbeitszeiten und vermindert um seine Minderarbeitszeiten. Die *produktiv nutzbare* Arbeitszeit hingegen bereinigt die reale Arbeitszeit um die durchschnittlichen (geplanten und ungeplanten) Fehlzeiten und die bezahlten Pausen, d.h. Kurzpausen. Unproduktiv sind sowohl die geplanten Fehlzeiten (z. B. Urlaub, Schulung) als auch die ungeplanten Fehlzeiten (z. B. Krankheit). Der Gesamtumfang des Urlaubs eines Agenten wird im Arbeitsvertrag vereinbart und steht somit fest. Die zeitliche Verteilung der Urlaubszeit auf den Planungszeitraum ist im Rahmen der gesetzlichen Vorgaben beeinflussbar.[122] Der Umfang von Schulungsmaßnahmen unterliegt keinen Bestimmungen. Hier sind die betrieblichen Regelungen zu berücksichtigen. Die ungeplanten Fehlzeiten unterliegen ebenfalls jahreszeitlichen Schwankungen. Sowohl die geplanten als auch die ungeplanten Fehlzeiten aber auch das Einplanen von Mehr- bzw. Minderarbeitszeiten führen dazu, dass die produktiv nutzbare Arbeitszeit im Planungszeitraum variiert. Die Summe der produktiv nutzbaren Arbeitszeiten aller Arbeitnehmer eines betrachteten Zeit-

[122] Vgl. dazu die Ausführungen in den Abschnitten 2.5.3.3.

raums bildet das Personalangebot in diesem Zeitraum.

	nominelle Arbeitszeit
+	Mehrarbeitszeit
-	Minderarbeitszeit
=	reale Arbeitszeit
-	geplante und ungeplante Fehlzeiten
-	Zeiten für die Bildschirmpausen
=	produktiv nutzbare Arbeitszeit des Agenten

Abbildung 3.1: Zusammenhang zwischen der nominellen, der realen und der produktiv nutzbaren Arbeitszeit eines Agenten

Bei der mittelfristigen Personalbedarfsbestimmung wird davon ausgegangen, dass alle im Call Center beschäftigten Arbeitnehmer eine identische nominelle Arbeitszeit haben. In diesem Kapitel wird für alle Agenten ein Vollzeitarbeitsverhältnis unterstellt. Darüber hinaus wird zunächst auf das Einplanen von Mehr- und Minderarbeitszeiten verzichtet.

In dem nächsten Abschnitt findet die Betrachtung der Determinante des Anrufvolumens statt. Die Vorhersage der Nachfrage ist entscheidend für die Bestimmung des Personalbedarfs, so dass die Möglichkeiten, die Nachfrage zu prognostizieren, untersucht werden. Auf eine detaillierte Beschreibung der Prognoseverfahren wird dabei verzichtet, da der Personalaspekt im Vordergrund dieser Arbeit steht. Demnach erfolgt im Abschnitt 3.2 ein Überblick über die bisherigen Anwendungen unterschiedlicher Prognoseverfahren im Call Center.

3.2 Vorhersage der Nachfrage

Die Vorhersage der Nachfrage (=Anrufvolumen) bildet die Grundlage für die Personalbedarfsermittlung. Eine schlechte Prognose kann zur Folge haben, dass die Anzahl der beschäftigten bzw. eingesetzten Agenten in den einzelnen Perioden über- bzw. unterdimensioniert ist. Dies kann dazu führen, dass entweder auf Kosteneinsparungen bzw. Gewinnzuwächse verzichtet wird oder aber die Zufriedenheit der Kunden mit dem angebotenen Service leidet. Klungle und Maluchnik zeigen beispielsweise anhand von AAA Michigan, einem Automobilclub in den USA, dass eine Verminderung des Vorhersagefehlers um 5 % eine Kostenersparnis von $100.000 pro Jahr in jedem Call Center des Unternehmens nach sich zieht.[123] Somit ist es für ein Call Center wichtig, ein

[123] Vgl. Klungle/ Maluchnik (1997), S. 10.

Prognoseverfahren auszuwählen, mit dem das Anrufvolumen möglichst gut abgebildet werden kann. Je nach Datenlage und kausalen Faktoren sind unterschiedliche Verfahren anwendbar. Welche der nachfolgend genannten Methoden die besten Prognoseergebnisse erreichen, hängt von den Gegebenheiten des jeweiligen Call Centers ab. Die in Frage kommenden Prognoseverfahren sollten anhand der Prognosegenauigkeit ausgewählt werden. Die Beurteilung der Eignung eines Prognoseverfahrens ist nicht Call Center spezifisch. Im Rahmen dieser Arbeit wird dazu auf die weiterführende Literatur verwiesen.[124] Der Prognosezeitraum sowie die Detailliertheit der Prognose hängen von der Planungsebene ab, die Grundprinzipien und die Verfahren, die generell zur Anwendung kommen können, sind jedoch identisch. Im Folgenden wird auf die Methoden der Zeitreihenmodelle und der kausalen Verfahren sowie auf den einfachen, in der Praxis häufig verwendeten Ansatz der top-down Planung eingegangen.

Call Center haben unter Zuhilfenahme der ACD-Anlage einen einfachen Zugriff auf historische Daten, unter anderem auf das Anrufvolumen vergangener Perioden.[125] Aus diesem Grunde wird bei der Prognose der Nachfrage häufig auf *Zeitreihenmodelle* zurückgegriffen, die das zukünftige Anrufvolumen basierend auf den vergangenen Anruf-zahlen prognostizieren, ohne dabei weitere kausale Einflussfaktoren zu berücksichtigen. Zu den für Call Center in Frage kommenden Zeitreihenmodellen zählen die Verfahren der Mittelwertbildung, der Grundansatz der exponentiellen Glättung sowie dessen Erweiterungen hinsichtlich saisonalen und trendförmigen Schwankungen und die ARIMA-Verfahren von Box und Jenkins.[126]

In der Literatur wird häufig die Prognose des Anrufvolumens eines Call Centers mittels Zeitreihenmodellen vorgestellt. Lin et al. bestimmen das stündliche Anrufvolumen mittels eines gleitenden Durchschnitts der letzten drei Monate.[127] Dabei betrachten sie unterschiedliche Zeitreihen für die einzelnen Wochentage und Stunden des Tages. Die Anwendung der exponentiellen Glättung wird von Klungle und Maluchnik aufgezeigt.[128] Sie verwenden die exponentielle Glättung für die langfristige Prognose des jährlichen und

---

[124] Vgl. z. B. DeLurgio (1998), Makridakis/ Wheelwright/ Hyndman (1998) und Gaynor/ Kirkpatrick (1994).

[125] Diese Daten müssen noch um die wiederholt anrufenden Kunden bereinigt werden, die beim ersten Anruf nicht bedient wurden.

[126] Zur Methodik der genannten Verfahren vgl. beispielsweise Makridakis/ Wheelwright/ Hyndman (1998), Hansmann (1995), Box/ Jenkins/ Reinsel (1994), Mertens (1994) sowie Hüttner (1986).

[127] Vgl. Lin /Lai/ Hung (2000).

[128] Vgl. Klungle/ Maluchnik (1997).

monatlichen Anrufaufkommens.

ARIMA-Verfahren können sowohl bei der Zeitreihenanalyse als auch bei der Prognose Anwendung finden.[129] Die möglichen ARIMA-Verfahren kombinieren autoregressive (AR) und integrierende (I) Bestandteile sowie gleitende Durchschnitte (moving average: MA). Die Gewichtung der Vergangenheitsdaten wird bei den ARIMA-Verfahren optimiert, indem durch den Einsatz von Filtern Instationaritäten der Zeitreihe eliminiert werden. Instationaritäten bezeichnen systematische Veränderungen wie Trends, Saisonalitäten, zunehmende Fluktuation der Daten und zufällig oder einmalig auftretende Ereignisse.[130] Die ARIMA-Verfahren bieten demnach die Möglichkeit, stationäre, nichtstationäre und saisonbehaftete Zeitreihen zu untersuchen. Gleichermaßen können diese Verfahren bekannte Ereignisse (Interventionen) wie Feiertage oder Marketingkampagnen sowie Ausreißer berücksichtigen.[131] ARIMA-Modelle werden in der Literatur als geeignete Verfahren für die Vorhersage des Anrufvolumens präsentiert, wobei verschiedene Autoren unterschiedlich lange Zeitintervalle prognostizieren.[132] Bianchi et al. vergleichen bei der Prognose des täglichen Anrufvolumens die Ergebnisse der exponentiellen Glättung mit denen des ARIMA-Verfahrens mit Intervention.[133] Sie gelangen zu dem Schluss, dass die Vorhersagegenauigkeit des ARIMA-Verfahrens besser ist als die der exponentiellen Glättung. Trotz der häufig berichteten guten Ergebnisse gestaltet sich ihr Einsatz aufgrund des umfangreichen mathematischen Hintergrundes in der Praxis als schwierig. Erweist sich die Verbesserung der Vorhersagegenauigkeit der ARIMA-Verfahren im Vergleich zu anderen Vorhersageverfahren als gering, wird im Hinblick auf den erheblichen Aufwand eher auf den Einsatz von ARIMA-Verfahren verzichtet.[134]

In einer umfassenden Untersuchung von Hauenschild wird anhand der Daten des Call Centers der Deutschen Telegate AG das halbstündliche Anrufvolumen mehrerer Verfahren

---

[129] Das Verfahren wird nur in seinem Grundgedanken dargestellt. Zur genauen Beschreibung siehe DeLurgio (1998) und Box/ Jenkins/ Reinsel (1994).

[130] Vgl. Hüttner (1986), S. 11.

[131] Antipov und Meade liefern einen Literaturüberblick zur Berücksichtigung von Marketingkampagnen in Prognoseverfahren. Vgl. Antipov/ Meade (2002).

[132] Vgl. Nijdam (1990) für die Vorhersage des monatlichen Anrufvolumens; Bianchi/ Jarrett/ Hanumara (1998), Andrews/ Cunningham (1995) und Buffa/ Cosgrove/ Luce (1976) für ein tägliches Prognoseintervall und Hauenschild (2000) für die halbstündliche Prognose.

[133] Vgl. Bianchi/ Jarrett/ Hanumara (1998).

[134] Vgl. Helber/ Stolletz (2001), S. 10 f.; Hauenschild (2000), S. 120 und Klungle/ Maluchnik (1997), S. 13.

miteinander verglichen.[135] Sie betrachtet dabei die Mittelwertbildung, die exponentielle Glättung erster und zweiter Ordnung, ARIMA-Modelle sowie neuronale Netze. Dabei wird die Datenreihe – ausgenommen im Falle der ARIMA-Modelle – in Tagtypen (z.B. Montag bis Sonntag, Feiertag in der Woche bzw. am Wochenende) und Halbstundentypen (z.B. Montag 8.30-9.00 Uhr) zerlegt. Für jede separate Zeitreihe eines Tagtyps und eines Halbstundentyps wird eine Prognose durchgeführt. Hauenschild zeigt, dass die ARIMA-Modelle, die explizit Feiertage berücksichtigen, das beste Ergebnis hinsichtlich der Wurzel des mittleren quadratischen Fehlers haben. Wird hingegen die Vorhersage mit Mittelwerten durchgeführt, sind die Prognosefehler am größten. Die Anwendung eines neuronalen Netzwerkes ergab im Vergleich ebenfalls keine zufriedenstellenden Ergebnisse. Die Methoden der exponentiellen Glättung liefern mit erheblich geringerem Aufwand als die ARIMA-Modelle gute Ergebnisse und stellen damit für die Praxis eine Alternative dar.

Sofern andere kausale Faktoren wesentlich das Nachfragevolumen beeinflussen, gestaltet sich der Einsatz von Zeitreihenmodellen oft ungenügend. In diesem Fall empfiehlt sich die Anwendung von *kausalen Verfahren*. Die Regressionsanalyse bietet sich in diesem Rahmen an. Mabert vergleicht mehrere Verfahren für die Prognose des täglichen Anrufvolumens und kommt zu dem Schluss, dass für seine Daten Zeitreihenmodelle nicht geeignet sind.[136] Er empfiehlt die Regressionsanalyse. Klungle und Maluchnik erhalten ebenfalls bei der Prognose des täglichen Anrufvolumens bei AAA gute Ergebnisse mit der Regressionsanalyse.[137] Ihre Daten hängen stark von den Temperaturen, der Anzahl der Mitglieder bei AAA sowie dem Beschäftigungsindex ab.

Die in der Literatur beschriebenen Verfahren für Call Center sagen – mit unterschiedlichen Methoden – überwiegend das tägliche Anrufvolumen vorher. Das prognostizierte tägliche Anrufvolumen lässt sich für Zwecke der mittelfristigen Planung leicht aggregieren. Die für operative Zwecke benötigten detaillierten Werte des Anrufvolumens werden häufig über eine prozentuale Aufteilung erreicht. Dieser Ansatz, der in der Praxis häufig durchgeführt und akzeptiert wird, basiert auf einer *top-down Planung*.[138] Beispielhaft wird diese beginnend bei der Ebene des monatlichen Anrufvolumens beschrieben, sie kann jedoch auch bei einer Ebene beginnen, deren Anrufvolumen einen kürzeren Zeitraum umfasst.

---

[135] Vgl. Hauenschild (2000), S. 117-120.

[136] Vgl. Mabert (1985).

[137] Vgl. Klungle/ Maluchnik (1997).

[138] Vgl. Gans/ Koole/ Mandelbaum (2003), S. 96 und Cleveland/ Mayben/ Greff (1998), S. 69-71.

Basierend auf den vergangenen monatlichen Anrufvolumina wird mittels eines Zeitreihenmodells das monatliche Anrufvolumen prognostiziert. Die wöchentlichen Anruferzahlen werden anhand von durchschnittlichen Anteilen der jeweiligen Woche und des Wochentages ermittelt. Sofern in eine Woche ein Feiertag fällt, werden separate Indexfaktoren für diese Woche erstellt, abhängig vom Wochentag des Feiertages. Die für die operative Planung notwendige Prognose des Anrufaufkommens für kleinere Zeitabschnitte eines Tages werden ebenfalls durch eine prozentuale Verteilung des ermittelten Tageswertes auf die entsprechenden Intervalle erreicht.[139] Die Prozentangaben ergeben sich aus den durchschnittlichen Anteilen, die das Anrufvolumen des Intervalls in der Vergangenheit am jeweiligen Wochentag erlangte. Buffa et al. teilen das tägliche Anrufvolumen mit Hilfe von exponentiell geglätteten Prozentsätzen auf die Halbstundenintervalle des Tages auf.[140] Wurden beispielsweise für einen Montag von 8:30-9:00 Uhr durchschnittlich 6,5 % der Anrufe des Tages ermittelt, so kann das Anrufaufkommen in diesem Zeitraum mit Hilfe des Prozentsatzes aus dem täglichen Anrufvolumen berechnet werden. Hauenschild stellt jedoch fest, dass eine detaillierte Vorhersage auf Basis von Halbstundenwerten zu besseren Ergebnissen führt als die top-down Planung, bei der sich die Halbstundenwerte aus den prozentualen Anteilen der Tageswerte ergeben.[141]

3.3 Anforderungen an die Methoden der Personalbedarfsermittlung

Bisher wurden die Nachfrage, die Bearbeitungszeiten und im Falle der mittelfristigen Personalbedarfsermittlung die Arbeitszeit als Determinanten des Personalbedarfs identifiziert. Während die Bearbeitungs- und Arbeitszeiten als gegeben angenommen werden können, sind die Anforderungen des Marktes über die Prognose der Nachfrage zu implementieren. Daher wurden im vorherigen Abschnitt geeignete Prognosemethoden der Nachfrage vorgestellt. In diesem Abschnitt findet die Erörterung der Anforderungen statt, denen die Methoden der Personalbedarfsermittlung speziell für Call Center unterliegen.

Zunächst ist zu klären, welche Anforderungen an die Methoden der Personalbedarfsermittlung im Call Center zu stellen sind. Der Bedarf an Agenten zur Beantwortung der Kundenanfragen muss sich an den Zielsetzungen des Call Centers orientieren.[142] Das bedeutet, dass bei der Bedarfsermittlung sowohl die angestrebten Maßzahlen der Kundenzufrieden-

---

[139] Vgl. beispielsweise Mabert (1985).

[140] Vgl. Buffa/ Cosgrove/ Luce (1976), S. 624.

[141] Vgl. Hauenschild (2000), S. 120.

[142] Die Zielsetzungen wurden im Abschnitt 2.4 beschrieben.

heit als auch die Kostenaspekte zu berücksichtigen sind. Um die Maßzahlen der Kundenzufriedenheit korrekt ermitteln zu können, ist es notwendig, das Call Center nicht als deterministisches System zu betrachten. Vielmehr ist sowohl die Stochastik der Zwischenankunftszeiten der anrufenden Kunden als auch die der Bearbeitungszeiten zu berücksichtigen. Eine Kalkulation auf Basis von Durchschnittswerten für die Zwischenankunftszeiten und die Bearbeitungszeiten geht implizit von einem gleichmäßigen Anrufeingang und identischen Bearbeitungszeiten der Anrufer aus. Je größer jedoch die Schwankungen in den Bearbeitungs- und Zwischenankunftszeiten sind, desto größer werden – bei gegebener Anzahl an Agenten und gegebenem Mittelwert – nicht nur die Wartezeiten für die Kunden, sondern auch die Leerzeiten der Agenten. Demnach sind Methoden der Bedarfsermittlung für Call Center geeignet, die in geeigneter Weise die Stochastik berücksichtigen.

Weiterhin ist der Nachfrageverlauf bei der Bedarfsermittlung zu beachten, denn die mittlere Zwischenankunftszeit unterliegt der im Abschnitt 2.5.1 beschriebenen Dynamik. Eine durchschnittliche Betrachtung der Zwischenankunftszeiten mehrerer Intervalle mit unterschiedlichem Nachfragevolumen ist möglich, sofern eine Linearität zwischen dem prognostizierten Nachfragevolumen und der Anzahl der eingesetzten Agenten besteht. Im Call Center Bereich ist eine konstante Produktivität, die besagt, dass jeder zusätzlich eingesetzte Agent eine konstante Outputänderung zur Folge hat, jedoch nicht gegeben.[143] Demzufolge ist eine durchschnittliche Betrachtung höchst bedenklich. Wird dennoch von einer durchschnittlichen Zwischenankunftszeit über Perioden mit sehr unterschiedlichen Nachfragevolumina ausgegangen, so führt das dazu, dass in vielen Intervallen der Personalbedarf überschätzt, in anderen Intervallen hingegen unterschätzt wird. Dabei gleicht sich die Über- und die Unterschätzung nicht aus. Das bedeutet, dass es zweckmäßig ist, einen Tag in Abschnitte zu unterteilen, wobei sich innerhalb eines Abschnittes die Zwischenankunftszeiten annähernd gleich gestalten.

Somit sind Methoden bei der Personalbedarfsermittlung auszuwählen, die sowohl die Stochastik als auch die Dynamik und die Zielsetzungen des Call Centers angemessen berücksichtigen. Basierend auf den im Abschnitt 3.1 genannten Bestimmungsfaktoren des Personalbedarfs schließt sich in den folgenden Abschnitten eine Beschreibung der Verfahren an, mit deren Hilfe sich die Bezugsgrößen in einen Personalbedarf übersetzen

[143] Eine ausführliche Begründung dieses Argumentes kann erst nach der detaillierten Darstellung der Warteschlangenmodelle erfolgen. Für die Argumentation wird auf Abschnitt 3.4.2.5 verwiesen.

lassen. Im Abschnitt 3.4 werden einige Methoden zur Bestimmung des kurzfristigen Personalbedarfs untersucht und für Call Center Anwendungen beurteilt. Abschnitt 3.5 widmet sich der Ermittlung des mittelfristigen Personalbedarfs.

3.4 Methoden zur Ermittlung des kurzfristigen Personalbedarfs

Es existieren zahlreiche Methoden zur Personalbedarfsermittlung. Im Call Center Bereich wurden bisher Regressionsanalysen, Warteschlangenmodelle, Simulationen, Kennzahlenmethoden sowie Fluidapproximationen eingesetzt. Simulationen kommen häufig bei Call Centern mit komplexen Strukturen zur Anwendung, bei denen aufgrund der Betrachtung heterogener Agenten und Anrufer Skills-Based Routing eingesetzt wird. Diese komplexen Strukturen führen dazu, dass keine exakte Methode existiert, die sowohl mathematisch handhabbar ist als auch das reelle System abbildet, so dass von der Simulation Gebrauch gemacht wird.[144] Eine Simulation gestaltet sich aufgrund der vielfältigen Möglichkeiten der Parametervariation als sehr aufwendige Methode. Im Rahmen dieser Arbeit werden Call Center mit homogenen Agenten und Anrufern betrachtet, die keine komplexen Strukturen aufweisen. Somit wird die Simulation in diesem Zusammenhang nicht weiter verfolgt.

Die Anwendung von Kennzahlenmethoden setzt eine konstante Produktivität voraus. Darüber hinaus unterstellen die Methoden einen gleichmäßigen Anrufeingang sowie konstante Bearbeitungszeiten. Diese Voraussetzungen sind im Call Center nicht gegeben, so dass der Einsatz von Kennzahlenmethoden abgelehnt werden muss.[145]

Die Idee der Fluidapproximation besteht darin, zufällige, zeitabhängige und diskrete Prozesse durch deterministische, zeitabhängige und kontinuierliche Prozesse zu ersetzen.[146] Die Anzahl der Kunden wird durch den Erwartungswert des zufälligen Prozesses repräsentiert. Das bedeutet, dass die Kunden im System nicht als diskrete Individuen dargestellt werden, sondern aus einer kontinuierlichen Anzahl bestehen. Je mehr Kunden sich im System befinden, desto eher kann die diskrete Natur der Kunden vernachlässigt werden.[147] Dies ist insbesondere bei großen Call Centern der Fall. Aufgrund der leichten Erstellung der Fluidapproximation in Form von Differenzialgleichungen findet sie

---

[144] Vgl. Atlason/ Epelmann (2004), S. 334 f.; Koole/ Pot/ Talim (2003) und Brigandi et al. (1994).

[145] Dennoch werden in der Literatur zur Ermittlung des mittelfristigen Personalbedarfs Kennzahlen vorgeschlagen. Vgl. z. B. Gamm (2000), S. 24 f. und Hughes (1995), S. 88 f.

[146] Vgl. Henken (2006), S. 30.

[147] Vgl. Borst/ Mandelbaum/ Reimann (2004) und Whitt (2003).

insbesondere bei Call Centern mit heterogenen Kunden und Agenten Anwendung, die eine Vielzahl an Kunden- und Agentengruppen zu berücksichtigen haben. Demnach ist der Ansatz für große Call Centern mit unterschiedlichen Kunden- und Agentengruppen passend. Aufgrund der Betrachtung kleiner Call Center mit homogenen Kunden und Agenten in dieser Arbeit wird die Fluidapproximation nicht weiter ausgeführt.[148]

Darüber hinaus existieren weitere Methoden der Personalbedarfsermittlung, die aus den unterschiedlichsten Gründen für Call Center nicht in Betracht kommen. Einige vernachlässigen die stochastische Komponente (z.B. Stellenplanmethode) oder beruhen auf Erfahrungen und Intuitionen der Anwender (z.B. Schätzverfahren, Trendanalogie) und berücksichtigen demnach ebenfalls die Stochastik nicht in angemessener Form. Arbeitswissenschaftliche Methoden (z.B. REFA, MTM) versuchen, über den Zeitbedarf einer Arbeitseinheit auf die benötigten Arbeitskräfte zu schließen. Eine lineare Transformation des Zeitbedarfs in einen Personalbedarf ist aufgrund der variablen Produktivität nicht möglich. Auf die Darstellung dieser Methoden wird daher verzichtet.[149] Im Folgenden wird eruiert, inwieweit die Regressionsanalyse und die Warteschlangenmodelle in der Lage sind, den Personalbedarf in Call Centern zu bestimmen.

3.4.1 Regressionsanalyse

Die Regressionsanalyse erklärt den Personalbedarf in Abhängigkeit von einer bzw. mehreren Einflussgrößen, indem diese in eine mathematische Beziehung gesetzt werden. Ein in der Vergangenheit ermittelter Zusammenhang der Einflussgröße(n) des Personalbedarfs wird im Anschluss daran zur Prognose verwendet.[150] Für den mathematischen Hintergrund der Regressionsanalyse sei auf die weiterführende Literatur verwiesen.[151]

Im Call Center wurde eine lineare Regressionsanalyse zur kurzfristigen Personalbedarfsbestimmung von Lin et al. sowie Lin eingesetzt.[152] Bei diesen Veröffentlichungen besteht die Maßzahl der Kundenzufriedenheit in der Einhaltung eines bestimmten Prozentsatzes

---

[148] Die Personalbedarfsermittlung mittels Fluidapproximation wird z.B. bei Henken (2006), Whitt (2006a) und (2006b), Harrison/ Zeevi (2005), Jiménez/ Koole (2004) und Mandelbaum/ Massey (1995) durchgeführt.

[149] Zur weiterführenden Literatur sei hier auf Drumm (2005), Bokranz (2004), Oechsler (2000), Scholz (2000), Kossbiel (1992) und Winnes (1978) verwiesen.

[150] Die Regressionsanalyse wurde beispielsweise erfolgreich zur Personalbedarfsbestimmung von Schwestern im Krankenhaus angewendet. Vgl. Helmer/ Oppermann/ Suver (1980).

[151] Vgl. beispielsweise Makridakis/ Wheelwright/ Hyndman (1998).

[152] Vgl. Lin/ Lai/ Hung (2000) und Lin (1999).

für die Auflegerate der anrufenden Kunden. Die Dynamik des Anrufvolumens wird bei dem Ansatz berücksichtigt, indem das Anrufvolumen in mehrere Zeitreihen zerlegt wird, die unterschiedliche Stunden- und Wochentagskombinationen beinhalten. Die Autoren betrachten anhand der zugrunde liegenden Daten der vergangenen 3 Monate für jede Stunde und jeden Wochentag die Korrelation zwischen Auflegerate und Arbeitslast.[153] Ergibt sich für eine Stunden- und Wochentagskombination ein hoher Korrelationskoeffizient (R^2>0,7), wird für diese Kombination die Anzahl an benötigten Agenten über die Regressionsgleichung ermittelt. Ein Korrelationskoeffizient größer als 0,7 ist jedoch nur in 60 % der von ihnen betrachteten Regressionsgleichungen der Fall. Dies begründet sich in der Tatsache, dass ein annähernd gleichbleibendes Anrufvolumen für eine Stunden- und Wochentagskombination unterstellt wird. Sofern das Anrufvolumen in einer betrachteten Stunden- und Wochentagskombination in der Vergangenheit größeren Schwankungen unterlag, lässt die Tatsache einer zunehmenden Produktivität bei steigendem Anrufvolumen den Korrelationskoeffizienten sinken.[154] Das bedeutet, dass die Regressionsanalyse ein relativ konstantes Anrufvolumen im Verlauf einer Stunden- und Wochentagskombination voraussetzt. Schwankungen des Anrufvolumens werden vernachlässigt und durch eine durchschnittliche Betrachtung ersetzt. Demnach wird die Stochastik hinsichtlich einer zielgerechten Personalbedarfsermittlung nicht ausreichend berücksichtigt. Die Regressionsanalyse ist somit nur begrenzt zur Personalbedarfsbestimmung im Call Center einsetzbar.

3.4.2 Warteschlangenmodelle

Bei Call Centern kommen bei der Personalbedarfsermittlung Methoden zum Einsatz, die die stochastische Komponente einbeziehen. Warteschlangen stellen für Call Center eine typische Methodik dar, wie man die zufällige Natur des Problems in den Griff bekommen kann. Warteschlangenmodelle berücksichtigen die Zufälligkeit u.a. im Anrufeingang und den Bearbeitungszeiten. Ihr Einsatz erfolgt im Rahmen der operativen Planung für Call Center für kurze Zeitabschnitte, um die Anzahl der benötigten Agenten einer Periode bei gegebener Zielsetzung zu bestimmen.

Im folgenden Abschnitt 3.4.2.1 werden zunächst die charakteristischen Merkmale eines

---

[153] Vgl. Lin/ Lai/ Hung (2000), S. 991 f. Die Arbeitslast einer Stunden-/Wochentagskombination *i,h* berechnet sich folgendermaßen: $Arbeitslast_{i,h} = \frac{Anrufvolumen_{i,h} \text{ /60 Minuten}}{\text{Anzahl an Agenten/ durchschnittliche Bearbeitungszeit}_{i,h}}$.

[154] Die steigende Produktivität wird im Abschnitt 3.4.2.5 einer genauen Untersuchung unterzogen.

Warteschlangenmodells beschrieben. Es schließt sich im Abschnitt 3.4.2.2 die Auswahl eines Warteschlangenmodells an. Das Auswahlkriterium besteht darin, dass der Ablauf eines Call Centers mit homogenen Agenten und Anrufern gut abgebildet wird. Im Abschnitt 3.4.2.3 wird das ausgesuchte Warteschlangenmodell beschrieben, um im Anschluss im Abschnitt 3.4.2.4 zu zeigen, wie sich mit dessen Hilfe bei einem prognostizierten Anrufvolumen der Personalbedarf für eine bestimmte Zielsetzung für Intervalle mit konstanter Anrufrate ermitteln lässt. Des Weiteren erfolgen im Abschnitt 3.4.2.5 mehrere kurze Anwendungsbeispiele, die einige Erkenntnisse veranschaulichen und insbesondere eine konstante Produktivität im Call Center Bereich beispielhaft widerlegen.

3.4.2.1 Charakterisierung eines Warteschlangenmodells für Call Center mit homogenen Kunden und Agenten

Warteschlangenmodelle werden häufig als Mittel zur Bestimmung des operativen Personalbedarfs im Call Center eingesetzt. Ein Call Center kann schematisch als Warteschlangenmodell gesehen werden. Das Modell besteht aus Anrufern, Agenten und Warteschlangen für die Anrufer. Wie bereits in Abbildung 2.1 dargestellt, stehen den Anrufern *c* Agenten zur Verfügung. Sofern alle Agenten im Gespräch sind, besteht für die Kunden die Möglichkeit, zu warten. Wartende Kunden können das System wieder verlassen. Unter der Annahme homogener Kunden und homogener Agenten lässt sich ein solches Warteschlangenmodell vollständig anhand des Ankunftsprozesses und des Warteverhaltens der Anrufer, der Verteilung der Bearbeitungszeiten der Agenten sowie der Begrenzung des Warteraums beschreiben.[155]

a) Ankunftsprozess

Anrufer eines Call Centers entscheiden unabhängig von bereits wartenden oder bedienten Kunden zu welchem Zeitpunkt sie im Call Center anrufen. Ist es gleichermaßen wahrscheinlich, dass zu bestimmten Uhrzeiten mehr Kunden anrufen als zu anderen Uhrzeiten, so kann der Ankunftsprozess als zeitinhomogener Poissonprozess modelliert werden.[156] Aufgrund der schwankenden Ankunftsrate im Zeitablauf kann der Ankunftsprozess innerhalb eines Tages als nicht stationär angesehen werden. Ein üblicher Ansatz, mit dieser Nichtstationarität des Ankunftsprozesses umzugehen, besteht im SIPP (stationary inde-

[155] Stolletz unterscheidet vier grundlegende Charakteristiken des Warteschlangenmodells für homogene und heterogene Kunden und Agenten im Call Center: Kundenprofil, Agentencharakteristik, Routingpolitiken und Begrenzung des Warteraums. Vgl. Stolletz (2003), S. 21.

[156] Vgl. Brown et al. (2002), S. 9 f. oder Koole/ Mandelbaum (2001), S. 13.

pendent period by period)-Ansatz.[157] Bei dieser Approximation werden kürzere Zeitabschnitte mit einer stückweise konstanten Ankunftsrate zugrunde gelegt. Darüber hinaus wird jede Periode (meist mit einer Länge von 15, 30 oder 60 Minuten) des Tages als unabhängig von den anderen Perioden betrachtet, so dass sich der Ankunftsprozess der Kunden in dem isolierten Zeitintervall anhand eines homogenen Poissonprozesses charakterisieren lässt. Gleichzeitig erfolgt die Annahme, dass sich die Ankunftsrate innerhalb des Intervalls annähernd konstant verhält und das steady-state in jedem Zeitintervall erreicht wird. Bezüglich der Zuverlässigkeit des SIPP-Ansatzes wurde gezeigt, dass diese bei einer sinkenden Länge des betrachteten Zeitintervalls, bei einer steigenden Bearbeitungsrate und bei einem sinkenden relativen Wechsel der Ankunftsrate innerhalb eines Intervalls ansteigt.[158]

Sofern Kunden keinen Agenten im Call Center erreichen, besteht die Chance, dass sie zu einem späteren Zeitpunkt erneut anrufen. In der Regel tätigen Kunden ihren Wiederholungsanruf jedoch nicht in dem gleichen Intervall, sondern diese Anrufe verteilen sich üblicherweise über künftige Zeitintervalle.[159] Demnach erhöht sich das Anrufvolumen eines späteren Zeitintervalls. Aufgrund der oben beschriebenen isolierten Betrachtung der einzelnen Intervalle können Warteschlangenmodelle ohne Wiederholungsanrufer zum Einsatz kommen. Dementsprechend werden die Zwischenankunftszeiten aufeinanderfolgender Anrufe als exponentialverteilt mit der Rate λ angenommen.

b) Warteverhalten

Anrufer warten nicht immer, bis sie einen Gesprächspartner im Call Center erreichen. Sie legen auf, weil sie vor der Warteschlange zurückscheuen oder weil sie während des Wartens die Geduld verlieren. Sieht man davon ab, dass die Anrufer über die voraussichtliche Wartezeit informiert werden, so kann ein anrufender Kunde mit Wahrscheinlichkeit β vor dem Warten zurückscheuen. Entsprechend wird der Kunde mit der Wahrscheinlichkeit $1-\beta$ den Warteraum betreten.

Es wird angenommen, dass Kunden, die nicht sofort auflegen, bereit sind, bis zum Erhalt der Dienstleistung eine gewisse Zeit zu warten. Die tolerierte Wartezeit der einzelnen Kunden differiert. Man kann jedoch davon ausgehen, dass ein Kunde nach einem

---

[157] Vgl. Green/ Kolesar/ Soares (2001).

[158] Vgl. Green/ Kolesar/ Soares (2001), S. 552-556.

[159] Vgl. Andrews/ Parsons (1989), S. 4.

zufälligen Wartezeitlimit *T* auflegt, sofern er noch nicht bedient wurde. Meistens wird die individuelle Wartezeittoleranz als exponentialverteilt angenommen.[160] Im weiteren Verlauf der Arbeit sei der Mittelwert der Exponentialverteilung ν^{-1}.

c) Verteilung der Bearbeitungszeiten

Die Bearbeitungszeiten eines Anrufes umfassen die Gesprächs- und Nachbearbeitungszeiten. Aus Vereinfachungsgründen wird im weiteren Verlauf lediglich die aus diesen beiden Teilen aggregierte Bearbeitungszeit zugrunde gelegt.[161] Die Verteilung der Bearbeitungszeit wird unterschiedlich gesehen. Während einige Autoren die Ansicht vertreten, dass exponentialverteilte Bearbeitungszeiten einer guten Approximation der Verteilung der Bearbeitungszeiten entspricht, kommen andere zu dem Ergebnis, dass diese näherungsweise Betrachtung nur unzureichend ist.[162] In dieser Arbeit werden, wie zumeist üblich, aus Vereinfachungsgründen die Bearbeitungszeiten aufgrund der Homogenität der Agenten als identisch verteilt, der Exponentialverteilung folgend, angenommen. In diesem Fall sind die Markoveigenschaften eines gedächtnislosen Prozesses gegeben, d.h. die zukünftige Dauer eines Gespräches ist unabhängig von der bereits vergangenen Gesprächsdauer. Die Bearbeitungszeiten folgen einer Exponentialverteilung mit dem Mittelwert μ^{-1}.

d) Begrenzung des Warteraums

Der Warteraum nimmt die Kunden auf, die aufgrund fehlender freier Agenten auf einen Gesprächspartner warten. Die Begrenzung des Warteraums limitiert die Anzahl der Kunden *K*, die maximal im System sein können. Das bedeutet, dass Kunden ein Besetztzeichen erhalten, sofern zum Zeitpunkt ihres Anrufes bereits *K* Kunden im System sind.

---

[160] Vgl. beispielsweise Whitt (2004), S. 1452 oder Brown et al. (2002), S. 23.

[161] Vgl. Stolletz (2003), S. 25. Dies führt zwar zu einer Überschätzung der Anzahl an besetzten Leitungen, da die Telefonleitung während der Nachbearbeitung eines Anrufes nicht besetzt ist, eine exakte Analyse gestaltet sich jedoch als schwierig. Sofern Nachbearbeitungszeiten kurz sind, ist diese Annahme zu vertreten.

[162] Zu Ersteren zählt beispielsweise Koole. Vgl. Koole (2002a), S. 43. Mandelbaum, Sakov und Zeltyn beschreiben, dass die Approximation gut ist, sofern Mittelwert und Standardabweichung der Bearbeitungszeiten nahe beieinander liegen. Vgl. Mandelbaum/ Sakov/ Zeltyn (2001), S. 44. Brown et al. zeigen hingegen für einen Datensatz, dass die Bearbeitungszeiten lognormalverteilt sind. Vgl. Brown et al. (2002), S. 17.

3.4.2.2 Auswahl eines geeigneten Warteschlangenmodells

In den meisten Call Center Anwendungen bedient man sich eines $M/M/c/\infty$-Warteschlangenmodells. Die Bezeichnung eines (.)/(.)/c/K-Modells geht auf Kendall zurück.[163] Das (.)/(.)/c/K-Modell beschreibt ein Warteschlangensystem, bei dem der erste Eintrag die Verteilung des Ankunftsprozesses und der zweite Eintrag die Verteilung der Bearbeitungszeiten charakterisiert. c konkretisiert die Anzahl der identischen Server. K bezeichnet die Anzahl an Kunden, die maximal im System sind. Wird ferner die Ungeduld der Anrufer berücksichtigt, so kann ein Zusatz ‚+(.)' nach dem letzten Eintrag die Wartezeittoleranz beschreiben. Das Symbol M steht für eine unabhängig identisch exponentialverteilte Zufallsvariable. Andere Symbole werden für weitere Verteilungen benutzt.

Das $M/M/c/\infty$-Modell wird in Call Center Kreisen als Erlang-C-Modell bezeichnet.[164] Bei diesem Warteschlangenmodell wird der Warteraum als unbeschränkt angenommen. Das bedeutet, dass Anrufer niemals das Besetztzeichen erhalten. Weiterhin wird eine unendlich große Geduld der Anrufer unterstellt. Demnach legen die Anrufer weder auf, noch scheuen sie vor der Warteschlange zurück. Aufgrund der Modellannahmen wird bei Anwendung dieses Warteschlangenmodells die Anzahl der benötigten Agenten häufig überschätzt, da die Anzahl der Kunden im System überschätzt wird. Dennoch wird das $M/M/c/\infty$-Modell aufgrund seiner Einfachheit (es werden nur die Parameter λ, μ und die Anzahl der Agenten c benötigt) in der Praxis häufig angewendet, insbesondere kommerzielle Softwareprogramme greifen darauf zurück.[165] Bei eigens für Call Center entwickelten Softwareprogrammen wird die Berücksichtigung der Ungeduld über nicht näher erläuterte „einfache Erweiterungen" vorgenommen, teilweise wird die Überschätzung der zu bedienenden Kunden mit Hilfe eines Abschlagsfaktors korrigiert.[166]

Es existiert eine Vielzahl von Warteschlangenmodellen, die es ermöglichen, das System Call Center realistisch abzubilden. Diese Modelle berücksichtigen die Leitungskapazitäten ebenso wie die Ungeduld der Anrufer. Darüber hinaus werden Warteschlangenmodelle beschrieben, die für Call Center mit heterogenen Agenten und Anrufern eingesetzt werden

---

[163] Vgl. Kendall (1953), S. 339-341.

[164] A.K. Erlang entwickelte zu Beginn des 20. Jahrhunderts Formeln zur Berechnung einiger Maßzahlen der Kundenzufriedenheit des $M/M/c/\infty$-Modells. Vgl. Brockmeyer et al. (1960), S. 138-155.

[165] Siehe beispielsweise Rahmenbetriebsvereinbarung INVISION (2003) im Communication Center INVISION § 7 oder Cleveland/ Mayben/ Greff (1998), S. 90.

[166] Vgl. Schümann (2003), S. 34 oder Fukunaga et al. (2002), S. 3.

können.[167] Brown et al. zeigen anhand des einfachsten Warteschlangenmodells mit Berücksichtigung von Auflegern, dem Erlang-A-Modell (*M*/*M*/*c*+*M*), dass solch relativ einfache Warteschlangenmodelle erstaunlich robust reagieren, auch wenn die theoretischen Voraussetzungen nicht mit den empirischen Daten übereinstimmen.[168] Zur Untersuchung der hier behandelten Problemstellung von kleinen Call Centern mit homogenen Agenten und Anrufern unter Berücksichtigung der Ungeduld der Anrufer wird versucht, mit Hilfe des *M*/*M*/*c*/*K*+*M*-Modells die Call Center Umgebung realistisch abzubilden. Dieses Modell wurde ausgewählt, weil dabei in Erweiterung des Erlang-C-Modells sowohl eine exponentialverteilte Wartezeittoleranz unterstellt wird als auch das Zurückscheuen vor der Warteschlange Berücksichtigung findet. Im Vergleich zu dem Erlang-A- bzw. Erlang-C-Modell erfolgt eine Begrenzung des Warteraums. Das *M*/*M*/*c*/*K*+*M*-Warteschlangenmodell wird im folgenden Abschnitt vorgestellt.

3.4.2.3 Beschreibung des *M*/*M*/*c*/*K*+*M*-Warteschlangenmodells

Die im Abschnitt 3.4.2.1 getroffenen Annahmen der exponentialverteilten Zwischenankunftszeiten und Bearbeitungszeiten ermöglichen, die Anzahl an Anrufern im System mit Hilfe eines Geburts- und Sterbeprozesses zu modellieren.[169] Dieser wird für das *M*/*M*/*c*/*K*+*M*-Warteschlangenmodell in Abbildung 3.2 dargestellt. Nachfolgend werden die Zustände sowie die Übergangswahrscheinlichkeiten beschrieben. Der Zustand beschreibt die „Anzahl der Kunden im System“. Der Zustand *n* bedeutet, dass *n* Kunden im System sind. Die Kunden im System warten entweder oder werden bedient. Der Zustand *n* geht in den Zustand *n+1* über, sofern ein Kunde in das System eintritt. Dies geschieht mit der Geburtsrate bzw. Ankunftsrate λ_n. Der Übergang vom Zustand *n* nach *n-1* hingegen erfolgt, wenn ein Kunde das System verlässt. Die Rate des Übergangs in den benachbarten Zustand beträgt μ_n.

---

[167] Für einen Literaturüberblick der verschiedenen Warteschlangenmodelle, die bei Call Centern zum Einsatz kommen können, sei auf Stolletz verwiesen. Vgl. Stolletz (2003), S. 38-45.

[168] Vgl. Brown et al. (2002), S. 49-51. Garnett et al. beschreiben eine exakte Analyse des Erlang-A-Modells und zeigen die theoretische Validität der Approximation für große Erlang-A-Systeme. Vgl. Garnett/ Mandelbaum/ Reimann (2002).

[169] Vgl. Stolletz (2003), S. 49 sowie Gross/ Harris (1998), S. 45-47.

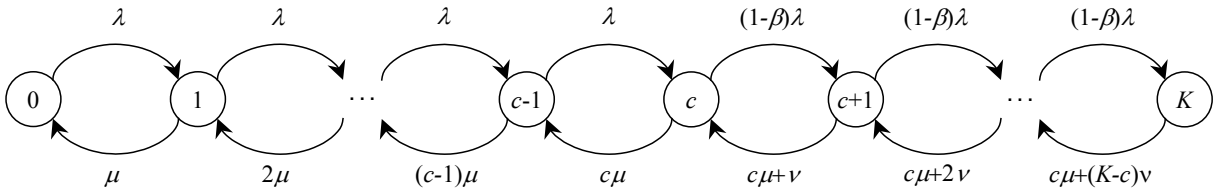

Abbildung 3.2: Übergänge des Geburts- und Sterbeprozesses

Im $M/M/c/K+M$-Warteschlangenmodell lassen sich die Geburts- und Sterberaten folgendermaßen beschreiben:[170]

$$\lambda_n = \begin{cases} \lambda, & \text{für } 0 \le n < c, \\ (1-\beta)\lambda, & \text{für } c \le n < K, \\ 0, & \text{für } n = K. \end{cases} \tag{3.1}$$

und

$$\mu_n = \begin{cases} n\mu & \text{für } 0 \le n \le c, \\ c\mu + (n-c)\nu, & \text{für } n > c. \end{cases} \tag{3.2}$$

Sofern kein ankommender Kunde wartet, erfolgt ein Zugang im Call Center mit der Rate λ. Sind alle Agenten belegt, dann wird die Ankunftsrate um den Anteil der Kunden β reduziert, die vor der Warteschlange zurückscheuen. Sind alle Leitungen belegt, dann werden die ankommenden Kunden blockiert, so dass kein weiterer Kunde Zugang zum Call Center erhält. Abgänge im Call Center finden statt, sobald ein bedienter Kunde das System verlässt. Ist die Anzahl der Kunden im System geringer als die Anzahl der Agenten, beträgt die Abgangsrate $n\mu$. Warten hingegen Kunden auf ihre Bedienung, so besteht die Abgangsrate nicht nur aus den bedienten Kunden, bei denen die Bedienrate $c\mu$ beträgt. Zusätzlich legt ein wartender Kunden mit der Rate ν auf. Die wartenden Kunden verlassen das System demnach mit der Rate $(n\text{-}c)\cdot\nu$.

Aus den Geburts- und Sterberaten lassen sich die Zustandswahrscheinlichkeiten p_n, dass sich n Anrufer im System befinden, berechnen:[171]

[170] Vgl. Whitt (1999), S. 195.

[171] Vgl. Stolletz (2003), S. 76.

$$p_n = \begin{cases} p_0 \dfrac{\lambda^n}{n!\,\mu^n}, & \text{für } 0 < n \le c, \\ p_0 \dfrac{(1-\beta)^{n-c}\,\lambda^n}{c!\,\mu^c \prod_{i=1}^{n-c}(c\mu + i\nu)}, & \text{für } c < n \le K, \\ \left(\sum_{n=0}^{c} \dfrac{\lambda^n}{n!\,\mu^n} + \sum_{n=c+1}^{K} \dfrac{(1-\beta)^{n-c}\,\lambda^n}{c!\,\mu^c \prod_{i=1}^{n-c}(c\mu + i\nu)} \right)^{-1}, & \text{für } n = 0. \end{cases} \tag{3.3}$$

Anhand der Zustandswahrscheinlichkeiten lassen sich einige Maßzahlen der Kundenzufriedenheit direkt ableiten. Die Wahrscheinlichkeit *P(B)* blockiert zu werden, lässt sich aus dem Zustand ableiten, dass ein Kunde im System ankommt, wenn alle Leitungen belegt sind. Das bedeutet für die Blockierwahrscheinlichkeit:

$$P(B) = p_K . \tag{3.4}$$

Befinden sich mehr als *c* Kunden im System und sind nicht alle Leitungen belegt, scheuen einige Anrufer vor der Warteschlange zurück. Die Wahrscheinlichkeit *P(Z)* vor der Warteschlange zurückzuscheuen, lässt sich berechnen durch:

$$P(Z) = \beta \sum_{n=c}^{K-1} p_n . \tag{3.5}$$

Die erwartete Warteschlangenlänge beträgt:

$$E[L] = \sum_{n=c+1}^{K} (n-c) p_n . \tag{3.6}$$

Sofern Kunden nicht vor der Warteschlange zurückscheuen, können sie während der Wartezeit auflegen. Die Wahrscheinlichkeit *P(AU)* aufzulegen, ergibt sich aus dem Verhältnis der Auflegerate der wartenden Anrufer zur Ankunftsrate.[172] Die Auflegerate lässt sich aus dem Produkt der erwarteten Warteschlangenlänge und der Rate *v*, mit der ein einzelner Anrufer auflegt, berechnen.

$$P(AU) = \frac{\nu \cdot E[L]}{\lambda} . \tag{3.7}$$

Wie bereits im Abschnitt 2.4.2.2 beschrieben, setzt sich die Bedienwahrscheinlichkeit *P(SV)* folgendermaßen zusammen: *P(SV)* = 1 – *P(B)* – *P(AU)* – *P(Z)*.

[172] Vgl. Stolletz (2003), S. 76 f.

Betrachtet man die Ankunftsrate der tatsächlich im Call Center ankommenden Kunden, so wird die Ankunftsrate λ um den Anteil der Kunden bereinigt, die das Call Center nicht erreichen, weil sie entweder blockiert werden oder vor der Warteschlange zurückscheuen. Dies spiegelt sich in der effektiven Ankunftsrate λ_{eff} wider. Für die effektive Ankunftsrate gilt:[173]

$$\lambda_{eff} = \sum_{n=0}^{c} p_n + (1-\beta)\lambda \sum_{n=c}^{K-1} p_n \,. \tag{3.8}$$

Mit Hilfe von λ_{eff} lassen sich die mittleren Wartezeiten für unterschiedliche Anrufergruppen ableiten.[174] Die mittlere Wartezeit der ankommenden Kunden beträgt:

$$E[WZ] = \frac{\sum_{n=c+1}^{K} (n-c)\, p_n}{\lambda_{eff}} \,. \tag{3.9}$$

Die mittlere Wartezeit der Aufleger ergibt sich folgendermaßen:

$$E[WZ \mid AU] = P(AU)^{-1} \sum_{n=c}^{K-1} p_n (1-\beta) \frac{\nu}{c\mu + (n-c+1)\nu} \sum_{j=1}^{n-c+1} \frac{j}{c\mu + j\nu} \,. \tag{3.10}$$

Die mittlere Wartezeit der bedienten Anrufer ergibt sich aus:

$$E[WZ \mid SV] = \frac{E[WZ]\,(P(SV) + P(AU)) - E[WZ \mid AU]\, P(AU)}{P(SV)} \,. \tag{3.11}$$

Zur Berechnung des Service Levels für verschiedene Anrufergruppen benötigt man die bedingte Wahrscheinlichkeit q_n, dass n Anrufer im System sind, unter der Bedingung, dass ein Anruf gerade ankommt. Für die Herleitung dieser Wahrscheinlichkeit sowie der nachfolgend beschriebenen Kriterien des Service Levels sei auf Stolletz verwiesen.[175]

$$q_n = \begin{cases} \dfrac{\lambda p_n}{\lambda_{eff}}, & \text{für } n < c, \\ \dfrac{(1-\beta)\lambda\, p_n}{\lambda_{eff}}, & \text{für } c \le n < K. \end{cases} \tag{3.12}$$

---

[173] Vgl. Stolletz (2003), S. 77.

[174] Vgl. Stolletz (2003), S. 77-79. Die unterschiedlichen Anrufergruppen wurden im Abschnitt 2.3 beschrieben.

[175] Vgl. Stolletz (2003), S. 79-82 und S. 174-177.

Die Wahrscheinlichkeit, innerhalb einer akzeptablen Wartezeit *t* bedient zu werden, beträgt für die bedienten Kunden:

$$P(WZ \le t|SV) = \frac{P(AU)+P(SV)}{P(SV)}\left(\sum_{n=0}^{c-1} q_n + \sum_{n=c}^{K-1} \frac{\prod_{i=0}^{n-c}(c\mu + i\nu)}{(n-c)!}\int_0^t \left(\frac{1-e^{-\varepsilon\nu}}{\nu}\right)^{n-c} e^{-\varepsilon(c\mu+\nu)}d\varepsilon\right). \quad (3.13)$$

Für die anrufenden Kunden lässt sich ebenfalls die Wahrscheinlichkeit berechnen, innerhalb einer akzeptablen Wartezeit *t* einen Agenten zu erreichen:

$$P(WZ \le t) = \left(\sum_{n=0}^{c-1} q_n + \sum_{n=c}^{K-1} q_n \frac{\prod_{i=0}^{n-c}(c\mu + i\nu)}{(n-c)!}\int_0^t \left(\frac{1-e^{-\varepsilon\nu}}{\nu}\right)^{n-c} e^{-\varepsilon(c\mu+\nu)}d\varepsilon\right)(1-P(B)-P(Z)). \quad (3.14)$$

Die mittlere Auslastung der Agenten lässt sich ermitteln durch:

$$E[U] = \sum_{n=1}^{c} p_n \frac{n}{c} + \sum_{n=c+1}^{K} p_n \,. \quad (3.15)$$

Die soeben dargestellten Erwartungswerte und Wahrscheinlichkeiten entsprechen den im Abschnitt 2.4.2.2 abgebildeten Maßzahlen der Kundenzufriedenheit. Diese können mit Hilfe des bei Stolletz beschriebenen Algorithmus berechnet werden.[176] Als Eingabeparameter werden die Anzahl der Agenten *c* sowie die Anzahl der zur Verfügung stehenden Leitungen *K*, die Ankunftsrate μ und die Bearbeitungsrate λ, die Auflegerate ν und der Parameter des Zurückscheuens β benötigt. Die Ankunftsrate ergibt sich aus dem Verhältnis von prognostiziertem Anrufvolumen und Intervalllänge. Die geschätzte Bearbeitungsrate lässt sich als Kehrwert der vergangenen mittleren Bearbeitungszeiten, die der ACD-Anlage entnommen werden können, ermitteln, sofern die Art und Weise der Gespräche keinen Änderungen unterliegt. Eventuell ist die Bearbeitungsrate zeitabhängig. Der Mittelwert der Wartezeittoleranz kann ebenfalls anhand der Daten der Telekommunikationsanlage geschätzt werden. Er ergibt sich aus:[177]

$$\nu^{-1} = \lambda^{-1}\frac{E[L]}{P(AU)}. \quad (3.16)$$

Der Anteil β der Kunden, die vor dem Warten zurückscheuen, kann geschätzt werden,

[176] Vgl. Stolletz (2003), S. 171-177.

[177] Vgl. Helber/ Stolletz (2004), S. 169 f.

indem aus den Daten der ACD-Anlage die Anzahl der wartenden Kunden der Anzahl an Kunden gegenübergestellt werden, die innerhalb einer Sekunde aufgelegt haben.

3.4.2.4 Personalbedarfsermittlung basierend auf dem Warteschlangenmodell

Zur Bestimmung der Anzahl der Agenten, die benötigt werden, um eine gewünschte Maßzahl der Kundenzufriedenheit zu erreichen, werden Warteschlangenmodelle häufig in Verbindung mit dem SIPP-Ansatz benutzt. Die Voraussetzungen des Einsatzes von Warteschlangenmodellen sind aber nur gegeben, wenn für den Ankunftsprozess Stationarität unterstellt werden kann und die Ankunftsrate innerhalb des betrachteten Zeitintervalls möglichst konstant ist. Aufgrund des schwankenden Anrufvolumens innerhalb eines Tages kann dies lediglich für Intervalle von kurzer Dauer angenommen werden. Dies ist bei Call Centern meist für eine Intervalllänge von 15, 30 oder 60 Minuten der Fall. Darüber hinaus muss man sich bewusst sein, dass sich die Zuverlässigkeit des SIPP-Ansatzes im Einzelfall unterschiedlich gestalten kann. Dies hängt von der Länge des gewählten Zeitintervalls, dem relativen Wechsel der Ankunftsrate innerhalb eines Intervalls und der Bedienrate ab.[178]

Für eine gegebene Ankunftsrate mit (unterstelltem) stationärem Ankunftsprozess und eine vorbestimmte Anzahl an Agenten können in einem kurzen Intervall mittels Warteschlangenmodell die vom Call Center ausgewählte(n) Maßzahl(en) der Kundenzufriedenheit berechnet werden. Die Ermittlung der Maßzahlen wurden im vorherigen Abschnitt dargestellt. Die übliche Vorgehensweise besteht darin, die kostenminimale (d.h. die geringst mögliche) Anzahl an Agenten c zu bestimmen, bei der eine bzw. mehrere Nebenbedingungen gelten, die die angestrebten Maßzahlen der Kundenzufriedenheit betreffen.[179] Die ausgewählten Maßzahlen werden meist anhand von Erfahrungen gewählt. Für die Maßzahlen wird üblicherweise eine obere bzw. untere Schranke definiert, die gerade noch als akzeptabel gilt. Die Anzahl freigeschalteter Wartepositionen spielt dabei nur eine geringe Rolle.[180]

---

[178] Vgl. Green/ Kolesar/ Soares (2001), S. 552-556 und die Ausführungen im Abschnitt 3.4.2.1.

[179] Für das $M/M/c/\infty$-Modell führen dies beispielsweise Brusco/ Jacobs (2001), Brusco et al. (1995), Andrews/ Parsons (1989), Gaballa/ Pearce (1979) und Segal (1974) durch.

[180] Vgl. Helber/ Stolletz (2004), S. 50 und Sparrow (1991), S. 170 f. Sparrow hebt für $M/M/c/K$-Modelle die Bedeutung der Bestimmung der Anzahl an Agenten im Vergleich zur Größe des Warteraums hervor. Denn die Erhöhung der Anzahl der Agenten verbessert sowohl die Erreichbarkeit des Systems als auch die Wartezeiten. Die Vergrößerung des Warteraums hingegen verbessert zwar die Erreichbarkeit, die Wartezeiten können sich aber erhöhen, da mehr Anrufer die Möglichkeit haben, den Warteraum zu betreten.

Methodisch geht man so vor, dass die Anzahl der Agenten in einer Periode so lange schrittweise erhöht wird, bis die angestrebte(n) Maßzahl(en) der Kundenzufriedenheit gerade überschritten wird bzw. werden. Diese Agentenzahl entspricht der kostenminimalen Anzahl, die das angestrebte Servicemaß erfüllt. Häufig besteht die Nebenbedingung für die Kundenzufriedenheit aus dem Service Level, bei dem mindestens 80 % der anrufenden Kunden innerhalb von 20 Sekunden bedient werden sollen. Gleichermaßen ist auch das Heranziehen anderer Maßzahlen möglich. Normalerweise ist die Schranke für die gewählte Nebenbedingung trotz zeitlich schwankender Nachfrage für alle Perioden identisch. Bei gewinnorientierten Ansätzen, die nicht die Kostenminimierung, sondern eine Gewinnmaximierung als Ziel anstreben, wird hingegen die Vorgabe eines fixen Service Levels aufgegeben. Hier schwankt der Service Level vielmehr als Funktion des variierenden Bedarfs. Die Ansätze zur Bestimmung der Anzahl an Agenten für gewinnorientierte Call Center sind der weiterführenden Literatur zu entnehmen.[181]

Helber et al. betrachten die Bestimmung der Anzahl an Agenten periodenübergreifend.[182] Sie untersuchen die Fragestellung, wie der Agentenbedarf über einen Planungszeitraum mit mehreren Perioden verteilt werden soll. Dabei erhöhen sie sukzessiv den Personalbedarf um einen Agenten in der Periode, bei dem das Verhältnis des Anstiegs der aggregierten Bedienwahrscheinlichkeit des betrachteten Planungszeitraums zu den zusätzlichen Kosten, bestehend aus Personal- und Telefonkosten, am größten ist. Zum Erhalt der aggregierten Größe wird die Bedienwahrscheinlichkeit einer Periode mit der relativen Ankunftsrate gewichtet. Somit erlangen Perioden mit hohem Anrufvolumen ein größeres Gewicht im Vergleich zu Perioden mit geringem Anrufvolumen. Die einzelnen Perioden haben keine fixe Vorgabe des Service Levels, die erreicht werden muss. Das Verfahren bricht ab, sobald die aggregierte Bedienwahrscheinlichkeit einen Mindestwert erreicht hat und die aggregierte Wartezeit der bedienten Anrufer einen maximal zulässigen Wert unterschritten hat.

Abgesehen von der direkten Berechnung der Maßzahlen der Kundenzufriedenheit für eine vorgegebene Anzahl an Agenten, kann die Personalbedarfsermittlung für Call Center anhand eines Wurzelprinzips erfolgen, das auf den Annahmen der Warteschlangenmodelle

---

[181] Hier sei beispielsweise auf Helber/ Stolletz/ Bothe (2005), Koole/ Pot (2004), Andrews/ Parsons (1993) und Quinn/ Andrews/ Parsons (1991) sowie auf den Überblick bei Stolletz (2003) verwiesen.

[182] Vgl. Helber/ Stolletz/ Bothe (2005), S. 13-30 und Helber/ Stolletz (2004), S. 171-184. Sie betrachten u.a. Call Center, bei der die Anzahl der einzusetzenden Agenten beschränkt ist.

basiert. Dabei wird die Anzahl an Agenten anhand der minimal benötigten Anzahl an Agenten ermittelt. Die minimale Anzahl beträgt λ/μ, sie wird jedoch um einen Sicherheitsfaktor erhöht, der proportional zu $(\lambda/\mu)^{1/2}$ ist. Borst et al. gewichten diese Wurzel mit einem Koeffizienten, der die Wartezeit- und Lohnkosten abwägt.[183] Demnach braucht keine explizite Maßzahl der Kundenzufriedenheit eingehalten werden, vielmehr wird versucht, simultan Personal- und Wartekosten zu minimieren. Allerdings betrachten sie das Wurzelprinzip für das $M/M/c/\infty$-Modell in großen Call Centern, d.h. sie berücksichtigen keine auflegenden, zurückscheuenden oder blockierten Kunden. Garnett et al. erweitern das Prinzip um Aufleger, während Aguir et al. das Verhalten der wiederholt anrufenden Kunden integrieren.[184] Die Autoren berichten, dass sie lediglich bei großen Call Centern gute Ergebnisse mit dem Wurzelprinzip erreicht haben. Der Wurzelansatz wird hier nicht weiter verfolgt, weil die Arbeit auf kleine Call Center beschränkt ist.

Welcher der beschriebenen Ansätze der Personalbedarfsermittlung letztlich zu empfehlen ist, hängt von dem jeweils betrachteten Call Center ab. Für kleine Call Center kommen insbesondere die periodenweise sukzessive Erhöhung der Anzahl der Agenten sowie der von Helber und Stolletz beschriebene Ansatz der periodenübergreifenden Personalbedarfsbestimmung in Betracht. Die Auswahl des zur Anwendung kommenden Ansatzes hängt insbesondere von den Zielsetzungen des Call Centers ab. Sofern alle Perioden der gleichen Gewichtung unterliegen und in jeder Periode die Einhaltung eines fixen Service Levels mindestens angestrebt wird, hätte der erste Ansatz der periodenweisen sukzessiven Erhöhung eine bevorzugte Anwendung verdient. Sollten hingegen Perioden mit hohem Anrufvolumen bevorzugt zur Betrachtung kommen, so ist der periodenübergreifende Ansatz zu wählen. Im weiteren Verlauf dieser Arbeit wird der beschriebene Ansatz praktiziert, bei dem periodenweise für die prognostizierte Anrufrate eine sukzessive Erhöhung der Anzahl der Agenten erfolgt, bis sich die angestrebten Maßzahlen der Kundenzufriedenheit ergeben. Im Folgenden findet die Veranschaulichung einiger grundlegender Zusammenhänge zwischen dem Anrufvolumen und der Anzahl benötigter Agenten statt, die für die Personalbestandsplanung vonnöten sind.

---

[183] Vgl. Borst/ Mandelbaum/ Reiman (2004).

[184] Vgl. Aguir et al. (2006) und Garnett/ Mandelbaum/ Reiman (2002).

3.4.2.5 Beispiele und Anwendungen

Anhand von einigen Beispielen soll gezeigt werden, wie das ausgewählte Warteschlangenmodell bezüglich unterschiedlicher Parametervariationen reagiert. Zunächst werden für das $M/M/c/K+M$-Warteschlangenmodell unter Variation der Agentenanzahl mehrere Maßzahlen der Kundenzufriedenheit mit Hilfe des bei Stolletz beschriebenen Algorithmus berechnet.[185] Dem Beispiel ist zugrunde gelegt, dass der Zeitraum, für den die Anrufrate als konstant gelten kann, 30 Minuten (= 1800 Sekunden) beträgt. Gleichermaßen wird ein konstantes Anrufvolumen von 100 Anrufen für das betrachtete Zeitintervall angenommen. Das entspricht einer Ankunftsrate von $\lambda = 100/1800$ pro Sekunde. Die gewählte Bearbeitungszeit beträgt im Mittel 80 Sekunden. Das resultiert in einer Bearbeitungsrate von $\mu = 1/80$. Ein Kunde scheut vor der Warteschlange mit einer Wahrscheinlichkeit von 5 % zurück, d.h. $\beta = 0{,}05$. Die durchschnittliche Wartezeit eines Kunden vor dem Auflegen beträgt 25 Sekunden. Daraus ergibt sich $v = 1/25$. Es wird eine konstante Anzahl an Leitungen von $K = 12$ unterstellt. Die sich aus diesen Parametern mittels Warteschlangenmodell ergebenden Maßzahlen der Kundenzufriedenheit werden in der Tabelle 3.1 veranschaulicht. Der Service Level umschreibt den Prozentsatz der bedienten bzw. aller anrufenden Kunden, die innerhalb von 20 Sekunden bedient werden.

		Service Level				erwartete Wartezeit		Wahrscheinlichkeit		
Agenten-anzahl	Anzahl an Warteplätzen	bedienter Kunden	anrufender Kunden	Erreich-barkeit	Auslastung der Agenten	ankommender Kunden	bedienter Kunden	des Blockierens	zurück-zuscheuen	aufzu-legen
c	*q = K-c*	*P(WZ ≤20|SV)*	*P(WZ ≤20)*	*P(SV)*	*E[U]*	*E[WZ]*	*E[WZ|SV]*	*P(B)*	*P(Z)*	*P(AU)*
1	11	67,1%	14,0%	20,9%	92,8%	19,53	16,87	0,00%	4,64%	74,48%
2	10	78,1%	31,3%	40,1%	89,0%	14,56	11,13	0,00%	4,07%	55,86%
3	9	86,2%	49,1%	56,9%	84,4%	10,27	7,24	0,00%	3,35%	39,70%
4	8	91,8%	65,2%	71,0%	78,9%	6,79	4,56	0,00%	2,55%	26,46%
5	7	95,5%	78,2%	81,9%	72,8%	4,16	2,73	0,00%	1,79%	16,36%
6	6	97,7%	87,5%	89,6%	66,3%	2,35	1,54	0,00%	1,14%	9,30%
7	5	98,9%	93,5%	94,5%	60,0%	1,21	0,81	0,01%	0,66%	4,82%
8	4	99,6%	96,9%	97,4%	54,1%	0,57	0,39	0,02%	0,34%	2,27%
9	3	99,8%	98,7%	98,8%	48,8%	0,24	0,17	0,05%	0,16%	0,95%

Tabelle 3.1: Mittels Warteschlangenmodell berechnete Maßzahlen der Kundenzufriedenheit bei einer Anrufrate von λ=100/1800

Anhand der Tabelle lässt sich – bei gleichbleibendem Anrufvolumen und steigender Anzahl an Agenten – der abnehmende Grenznutzen des Service Levels eines zusätzlichen Agenten erkennen. Mit der Anzahl der Agenten steigt auch die Erreichbarkeit des Call Centers, d.h. die Bedienwahrscheinlichkeit eines Anrufers. Die Auslastung der Agenten

[185] Vgl. Stolletz (2003), S. 171-177 sowie die Ausführungen im Abschnitt 3.4.2.3.

nimmt hingegen mit jedem zusätzlichen Agenten ab. Die erwartete Wartezeit sinkt ebenfalls mit der steigenden Anzahl der Agenten. Die Wahrscheinlichkeit des Blockierens von Anrufern ist aufgrund der hohen Leitungsanzahl bei einer geringen Anzahl an Agenten nicht vorhanden. Selbst bei einer geringen Anzahl an Warteplätzen ist die Wahrscheinlichkeit des Blockierens immer noch vernachlässigbar. Die Wahrscheinlichkeit des Zurückscheuens von Anrufern vor der Warteschlange und dem Auflegen während des Wartens nimmt mit steigender Agentenanzahl ab, da die Kunden im Schnitt weniger warten müssen.

Soll mittels Warteschlangenmodell der Personalbedarf für das Anrufvolumen von 100 prognostizierten Anrufern bestimmt werden, so muss die zugrunde liegende Zielsetzung beachtet werden. Besteht die Zielsetzung beispielsweise darin, mindestens 80 % der anrufenden Kunden innerhalb von 20 Sekunden zu bedienen, so ist die minimale Agentenanzahl zu wählen, die den Service Level von 80 % überschreitet. Demnach sind 6 Agenten erforderlich. Die entsprechende Zeile ist in der Tabelle 3.1 grau hinterlegt. Allerdings wird der angestrebte Service Level um 7,5 % überschritten. Agenten können jedoch nur im ganzen eingesetzt werden. Demzufolge ist entweder die Überschreitung des angestrebten Service Levels in Kauf zu nehmen, oder es ist möglich, vom angestrebten Service Level abzuweichen. Dies kann beispielsweise in Verbindung mit einem sekundären Zielkriterium entschieden werden. Alternativ ist denkbar, diese Entscheidung in Abhängigkeit vom Anrufvolumen der Periode zu treffen. Da die Zuwachsraten des Zielkriteriums beim Einsatz eines zusätzlichen Agenten um so größer sind, je geringer das Anrufvolumen in einer Periode ist, kann bei einem geringen Anrufvolumen ein Verzicht auf die vollständige Einhaltung des Service Levels unter Kostengesichtspunkten befürwortet werden. Es ist auch möglich, diese Entscheidung im Rahmen der Personaleinsatzplanung zu treffen.[186]

Der abnehmende Grenznutzen des Service Levels eines zusätzlichen Agenten legt mit steigendem Anrufvolumen eine zunehmende Produktivität der Agenten nahe. Dies wird beispielhaft für die bereits benutzten Parameter (Periodenlänge: 1800, $\mu = 1/80$, $\beta = 0{,}05$, $v = 1/25$) gezeigt. Bei der genannten Konstellation wird anhand des Warteschlangenmodells für unterschiedliche Anrufvolumina die minimale Anzahl an Agenten bestimmt, die benötigt werden, um den Service Level der anrufenden Kunden von 80/20 in der entsprechenden Periode zu realisieren. Da für dieses Beispiel bei den unterschiedlichen

[186] Dies wird im Abschnitt 4.1.2.3 erörtert.

Anrufvolumina die Anzahl der benötigten Agenten erheblich variiert, ist es sinnvoll, den Warteraum nicht für alle Konstellationen gleich groß zu gestalten. Dies begründet sich in der Tatsache, dass eine hohe Anzahl von Leitungen gerade bei einem geringen Anrufvolumen und somit bei einer geringen Anzahl benötigter Agenten einen unbegrenzten Warteraum impliziert. Um dies zu verhindern, wird der Warteraum unabhängig von der Anzahl der benötigten Agenten auf 10 Warteplätze begrenzt, d.h. $K = c + 10$. Die sich daraus ergebende Anzahl an Agenten sowie einige Maßzahlen der Kundenzufriedenheit sind im Anhang 9.1 zu finden. Setzt man für die verschiedenen Anrufvolumina die bedienten Kunden in Relation zu der Agentenanzahl, die für die Erreichung der Zielsetzung mindestens benötigt wird, so ergibt sich der durchschnittliche Ertrag eines Agenten. Der Zusammenhang zwischen Anrufvolumen und dem Durchschnittsertrag eines Agenten wird in Abbildung 3.3 veranschaulicht.[187] Der Durchschnittsertrag ist nicht konstant, sondern nimmt tendenziell mit der Erhöhung des Anrufvolumens bei gleichbleibender Intervalllänge zu. Das bedeutet bei Betrachtung des Produktionskoeffizienten, dem reziproken Wert des Durchschnittsertrages, dass dieser sich ebenfalls in Abhängigkeit vom Anrufvolumen ändert. Der Produktionskoeffizient nimmt mit der Erhöhung des Anrufvolumens ab. Betrachtet man den linken Teil der Abbildung 3.3, so erkennt man im Bereich des niedrigen Anrufvolumens eine stark zunehmende Anzahl bedienter Kunden je Agent, sofern das Anrufvolumen geringfügig gesteigert wird. Diese Situation eines geringen Anrufaufkommens ist insbesondere in kleinen Call Centern anzutreffen. Die steigende Anzahl bedienter Kunden pro Agent nähert sich mit steigendem Anrufvolumen einem Grenzwert, denn die Agenten können in dem betrachteten Intervall nicht mehr Zeit für Gespräche aufwenden, als ihre zur Verfügung gestellte Arbeitszeit. Der Wert, der einer 100%-igen Auslastung der Agenten entspricht, ist demnach ihre maximale Kapazität und stellt entsprechend den Grenzwert dar. Die „Sprünge“ der Kurve des Durchschnittsertrages kommen durch die diskrete Erhöhung der Agentenanzahl und der damit verbundenen sprunghaften Veränderung der einzelnen Zielkriterien zustande. Sofern das Anrufvolumen um einen Anruf gesteigert wird, die vorherige Anzahl an Agenten jedoch nicht ausreicht, um dieses Anrufvolumen zielgerecht zu erfüllen, ist ein zusätzlicher Agent einzuplanen. Im Vergleich zum Anrufvolumen, das um einen Anruf vermindert war, findet eine höhere Erreichung des zugrunde gelegten Zielkriteriums des Service Levels

[187] Ein vergleichbarer Zusammenhang ergibt sich, sofern man statt des Service Levels der bedienten Kunden den Service Level der anrufenden Kunden zugrunde legt.

statt. Die Anzahl der bedienten Kunden pro eingesetztem Agenten sinkt hingegen bei der Hinzunahme des Agenten, so dass der Durchschnittsertrag im Vergleich zu der um einen Anrufer verminderten Anzahl sinkt. Je größer das Anrufvolumen jedoch ist, desto geringer fallen die Sprünge aus und damit auch die Differenz der Zielkriterien beim Einsatz von einer Anzahl an Agenten, die sich um einen unterscheidet. Anhand dieser Grafik ist ersichtlich, dass große Call Center mit hohem Anrufvolumen economies of scale ausnutzen können, da bei ihnen eine hohe Auslastung der Agenten mit der gleichzeitigen Einhaltung eines angestrebten Service Levels einhergeht.

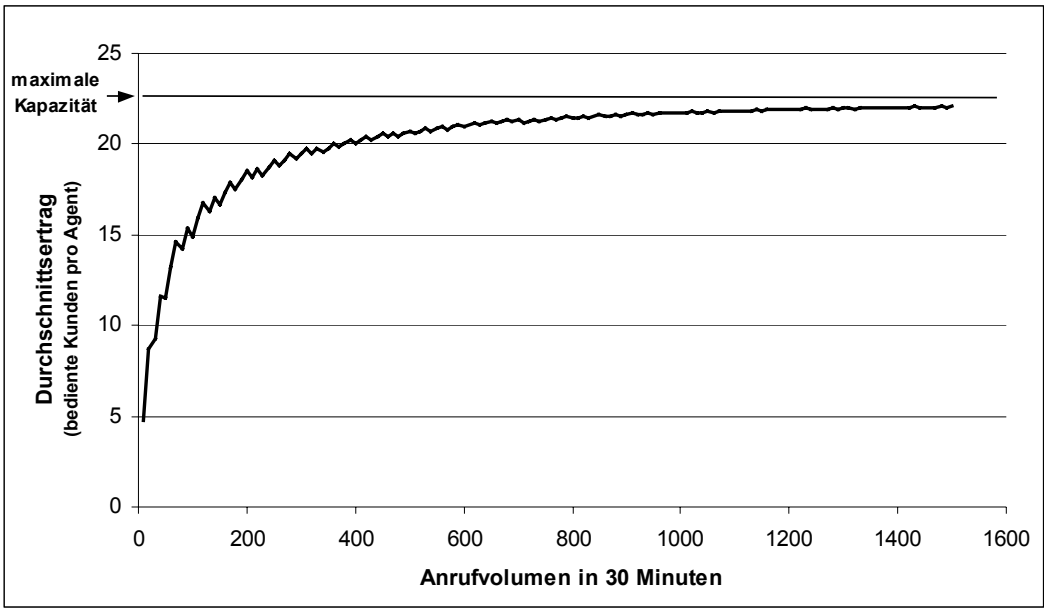

Abbildung 3.3: Zusammenhang zwischen Anrufvolumen und Durchschnittsertrag bei der minimalen Anzahl benötigter Agenten zur Erreichung des angestrebten Service Levels von 80/20

Eine tendenziell steigende Anzahl bedienter Kunden pro Agent bei wachsendem Anrufaufkommen bedeutet, dass eine Verdopplung des Anrufvolumens in einer Periode nicht zu einer Verdoppelung der benötigten Agenten führt. Vielmehr müssen weniger Agenten als die doppelte Anzahl eingesetzt werden.

Mittels eines Warteschlangemodells lässt sich der Personalbedarf eines Call Centers für Intervalle bestimmen, in denen die Ankunftsrate stationär ist. Dies kann für Intervalllängen von maximal 60 Minuten angenommen werden. Es bleibt zu klären, ob Warteschlangenmodelle zur Ermittlung des mittelfristigen Personalbedarfs angewendet werden können. Kleine Call Center haben in der Regel ein geringes Anrufvolumen, so dass sie insbesondere mit einer geringen Auslastung der Agenten und somit mit einem geringeren Durchschnittsertrag konfrontiert werden. Das bedeutet, dass eine Mittelwertbetrachtung des

Anrufvolumens über längere Zeitabschnitte bei starken halbstündlichen Schwankungen im Anrufvolumen erhebliche Implikationen hat. Dies resultiert aus der Tatsache, dass die Ankunftsrate erheblichen Schwankungen innerhalb des längeren Zeitabschnitts unterliegt und keine Stationarität unterstellt werden kann. Berechnet man dennoch den Personalbedarf mittels Warteschlangenmodell für ein Intervall mit nichtstationärer Ankunftsrate, indem man die mittlere Ankunftsrate als Eingangsparameter des Warteschlangenmodells benutzt, so hat das Auswirkungen bei der Personalbedarfsermittlung. Im Vergleich zur korrekten Berechnung des Personalbedarfs für kurze Intervalle mit stationärer Ankunftsrate, in denen die tatsächliche Ankunftsrate höher ist als die angenommene mittlere Ankunftsrate, wird ein zu geringer Personalbedarf berechnet. In Perioden hingegen, in denen die tatsächliche Ankunftsrate niedriger als die mittlere Ankunftsrate ist, übersteigt der berechnete Personalbedarf den benötigten. Die bereits gezeigten steigenden Durchschnittserträge führen dazu, dass – bei einer gegebenen Zielsetzung – die zu groß berechnete Anzahl an Agenten in den Perioden mit überdurchschnittlichem Anrufvolumen die fehlenden Agenten in den Perioden mit unterdurchschnittlichem Anrufvolumen nicht ausgleichen.

Dies kann beispielhaft anhand eines realen Anrufverlaufes eines kleinen Call Centers für einen Tag verdeutlicht werden. Das Call Center ist von 6:30 bis 22:00 Uhr, also 15,5 Stunden, betriebsbereit. Unter Verwendung der bereits benutzten Parameter ($\mu = 1/80$, $\beta = 0{,}05$, $v = 1/25$, $K = c+10$) soll die Anzahl der benötigten Agenten bestimmt werden, sofern in jeder Periode ein Service Level der anrufenden Kunden von 80/20 mindestens zu erfüllen ist. Unterteilt man die Öffnungszeit in Perioden mit 30-minütiger Länge, so ergeben sich mittels $M/M/c/K+M$-Modell die entsprechenden Agentenzahlen, die in Tabelle 3.2 dargestellt sind. In Summe werden für die 1408 Anrufe des Tages 107 Agentenhalbstunden benötigt. Wird hingegen eine durchschnittliche tägliche Ankunftsrate (λ=1408/(15,5·60·60)) zugrunde gelegt, so würden über den Tag verteilt bei der angestrebten Zielsetzung in jeder Periode 3 Agenten eingesetzt. Das bedeutet in Summe einen Agenteneinsatz von 93 Agentenhalbstunden an diesem Tag. Der Personalbedarf würde bei durchschnittlicher Betrachtungsweise erheblich unterschätzt, es würden 13,08 % zu wenig Agentenhalbstunden eingeplant.

		Periodenlänge von 30 Minuten		Periodenlänge von 15,5 Stunden	
Periodenbeginn	**Anrufvolumen**	**Personalbedarf**	**Service Level**	**Personalbedarf**	**Service Level**
06:30	5	1	83,8%	3	99,9%
07:00	7	2	97,2%	3	99,8%
07:30	28	3	92,3%	3	92,3%
08:00	35	3	88,0%	3	88,0%
08:30	53	4	89,0%	3	75,9%
09:00	60	4	85,5%	3	71,2%
09:30	72	5	89,6%	3	63,7%
10:00	71	5	90,0%	3	64,3%
10:30	72	5	89,6%	3	63,7%
11:00	75	5	88,5%	3	62,0%
11:30	100	6	87,5%	3	49,1%
12:00	79	5	86,9%	3	59,7%
12:30	65	4	83,0%	3	68,0%
13:00	73	5	89,3%	3	63,1%
13:30	65	4	83,0%	3	68,0%
14:00	75	5	88,5%	3	62,0%
14:30	75	5	88,5%	3	62,0%
15:00	70	4	80,4%	3	64,9%
15:30	55	4	88,0%	3	74,5%
16:00	58	4	86,5%	3	72,5%
16:30	40	3	86,9%	3	86,9%
17:00	42	3	83,3%	3	83,3%
17:30	32	3	89,9%	3	89,9%
18:00	25	3	93,9%	3	93,9%
18:30	23	2	81,8%	3	94,9%
19:00	15	2	90,1%	3	98,2%
19:30	20	2	85,0%	3	96,3%
20:00	9	2	95,7%	3	99,5%
20:30	7	2	97,2%	3	99,8%
21:00	1	1	96,3%	3	100,0%
21:30	1	1	96,3%	3	100,0%
Summe	1408	107	-	93	-
Mittelwert	45,42	3,45	88,8%	3,0	79,6%

Tabelle 3.2: Vergleich der Berechnung der Anzahl der Agenten und des Service Levels bei unterschiedlichen Periodenlängen

Insbesondere die Erreichung des Service Levels muss untersucht werden. Für die in Abhängigkeit von der Nachfrage berechneten Agentenzahlen wird das Zielkriterium in jedem Intervall mindestens erfüllt. Häufig, insbesondere in Perioden mit einer geringen

Nachfrage, wird das Kriterium stark übererfüllt, was zu einem durchschnittlichen Service Level von 88,8 % führt. Vergleicht man beispielsweise den Service Level für die Anrufvolumina der ersten beiden Intervalle, so genügt für das Anrufvolumen von 5 Anrufen ein Agent, um einen Service Level von 83,8 % zu erreichen. 7 Anrufer hingegen würden beim Einsatz eines Agenten einen Service Level von 78,5 % erhalten. Demnach ist das Zielkriterium von 80 % nicht erfüllt, so dass 2 Agenten einzuplanen sind. Der Service Level steigt sprunghaft auf 97,2 % an. Im Falle der Betrachtung einer durchschnittlichen täglichen Anrufrate hingegen wird in den beiden ersten Intervallen jeweils ein Personalbedarf von 3 Agenten zugrunde gelegt, was zu einer Zielerreichung nahe 100 % führt. Allerdings wird der durchschnittliche Service Level von 79,6 % dem Zielkriterium nicht gerecht, da in Perioden mit hohem Anrufvolumen der Service Level von 80/20 weit unterschritten wird. Das bedeutet, dass in Perioden mit hoher Nachfrage sehr viele Kunden länger als 20 Sekunden auf die Bearbeitung warten müssen, während in Perioden mit niedriger Nachfrage kaum ein Kunde diese Wartezeit übersteigt. Demnach rufen die überwiegende Mehrzahl der Kunden das Call Center zu Zeiten mit einem schlechten Service Level an, während wenige Kunden dies zu Zeiten mit einem guten Service Level tun. Diese Situation ist weder für die Kunden noch für die Agenten befriedigend.

Dies zeigt, dass für Perioden längerer Dauer mit nichtstationärer Ankunftsrate die direkte Anwendung des Warteschlangenmodells zu erheblichen Abweichungen zwischen berechneten und tatsächlich benötigten Agenten besteht. Somit lassen sich die Warteschlangenmodelle nicht direkt zur mittelfristigen Personalbedarfsermittlung einsetzen, da im Normalfall keine stationäre Ankunftsrate in dem Betrachtungszeitraum existiert.

3.5 Methode zur Ermittlung des mittelfristigen Personalbedarfs

Nach der Charakterisierung der Warteschlangenmodelle und der Vorstellung des $M/M/c/K+M$-Modells für Call Center mit homogenen Anrufern und Agenten im Abschnitt 3.4.2 erfolgte im weiteren Verlauf des Abschnitts die Darstellung, ob mit Hilfe des Warteschlangenmodells in Intervallen mit stationärer Anrufrate die Ermittlung des Personalbedarfs einer Periode möglich ist. Für Intervalllängen, für die keine stationäre Ankunftsrate unterstellt werden kann, führt die direkte Anwendung der Warteschlangenmodelle zu erheblichen Abweichungen bei der verfolgten Maßzahl der Kundenzufriedenheit. Demnach ist der mittelfristige Personalbedarf nicht anhand von Warteschlangenmodellen zu ermitteln. Im Folgenden wird untersucht, inwieweit sich Teilbereiche der Warteschlangenmodelle dennoch für die mittelfristige Personalbedarfsermittlung einsetzen lassen.

Warteschlangenmodelle lassen sich nicht direkt für die mittelfristige Personalbedarfsermittlung einsetzen, da keine stationäre Ankunftsrate in einem längeren Betrachtungszeitraum existiert. Man kann sich jedoch eines Hilfskonstruktes bedienen. Es wäre beispielsweise möglich, das Nachfragevolumen detailliert vorherzusagen, d.h. für Intervalle mit kurzer Länge, in denen die Ankunftsrate als stationär angenommen werden kann. Für diese Perioden lässt sich mittels Warteschlangenmodell ein Personalbedarf berechnen. Im Anschluss daran werden die Personalbedarfe der einzelnen Perioden für den betrachteten Zeitraum summiert. Das bedeutet, dass nicht das Anrufvolumen, sondern das bereits in einen Personalbedarf übersetzte Anrufvolumen einzelner Perioden des Planungshorizontes – mit jeweils unterstellter stationärer Ankunftsrate – zusammengefasst wird, um quantitative Aussagen bezüglich des Personalbedarfs eines Zeitraums treffen zu können, für den keine stationäre Ankunftsrate angenommen werden kann.

Somit lässt sich zusammenfassen, dass es notwendig ist, zunächst die Anruferzahlen der Perioden, in denen eine stationäre Anrufrate unterstellt werden kann, in einen kurzfristigen Personalbedarf zu transformieren, bevor der Personalbedarf über mehrere Perioden zusammengefasst werden kann, um auf den mittelfristigen Personalbedarf zu schließen, dem aufgrund der Länge des Planungshorizonts keine Stationarität zu unterstellen ist. Die Ermittlung des mittelfristigen Personalbedarfs bildet zwar nicht die Basis der Personalbestands- und Personaleinsatzplanung, zu Vergleichszwecken ist es jedoch notwendig zu zeigen, wie sich die Personalbestandsplanung von der mittelfristigen Personalbedarfsermittlung unterscheidet.

Im Folgenden wird ausgehend von dem vorab mittels Warteschlangenmodell berechneten kurzfristigen Personalbedarf, der im Anschluss daran für mehrere Perioden summiert wurde, eruiert, wie sich der mittelfristige Personalbedarf ermittelt lässt. Dies wird anhand einer Kennzahlenmethode durchgeführt. Kennzahlenmethoden basieren auf einer Bestimmungsgleichung, die die Arbeitsmenge ins Verhältnis zu Leistungsfähigkeit eines Mitarbeiters setzt. Aus dieser Bestimmungsgleichung lassen sich verschiedenen Formen von Modellen ableiten.[188] Sofern der mittels *M*/*M*/*c*/*K*+*M*-Modell berechnete und im Anschluss daran aggregierte kurzfristige Personalbedarf Eingang in eine Kennzahl findet, sind die Ankunftsrate und die Bearbeitungszeit mit deren Verteilung vorhanden.

[188] Vgl. Scholz (2000), S. 291-305; RKW (1996), S. 95 und Kossbiel (1992), Sp. 1603-1606.

Im Folgenden wird eine Kennzahlenmethode beschrieben, die aufgrund der Annahme der ausschließlichen Beschäftigung von Vollzeitagenten auch den Namen Vollzeitäquivalente trägt. Die Anzahl an Arbeitnehmern in Vollzeitbeschäftigung (*VZÄ*) für einen Planungszeitraum der Länge *T* beträgt:

$$\text{VZÄ} = \frac{\sum_{t=1}^{T} \text{Personalbedarfsstunden}_t \cdot (1 + \textit{Pufferzuschlag})}{\text{produktiv nutzbare Arbeitszeit eines Vollzeitagenten}} \tag{3.17}$$

Dabei beinhaltet der Zähler die anfallende Arbeitsmenge zuzüglich eines Zuschlages. Dieser Zuschlag ist notwendig, weil die Abdeckung der einzelnen kurzfristigen Personalbedarfsstunden nicht nur in Summe erforderlich ist. Vielmehr sind sie zu einem bestimmten Zeitpunkt zu befriedigen, nämlich beim Auftreten des entsprechenden Nachfragevolumens. Das bedeutet, dass sichergestellt sein muss, dass im Rahmen der Personaleinsatzplanung die wöchentliche Bedarfsstruktur der Personalbedarfsstunden durch die berechnete Anzahl der Vollzeitmitarbeiter tatsächlich abgedeckt werden kann. Im Ergebnis der Personaleinsatzplanung existieren insbesondere bei kleinen Call Centern häufig Perioden, in denen mehr Agenten zum Einsatz kommen als der benötigte Personalbedarf vorgibt. Denn aufgrund eines im Vergleich zu großen Call Centern geringen Personalbedarfs in den einzelnen Perioden werden insgesamt weniger Vollzeitschichten benötigt. Durch die geringe Anzahl an Vollzeitschichten gestaltet sich die Annäherung an ein volatiles Personalbedarfsmuster schwierig. Sofern der Personalbedarf in den einzelnen Perioden größer ist, hat der Personaleinsatzplaner einen größeren Spielraum bei der Auswahl an Schichten und deren Kombinationen, um ein bestehendes Personalbedarfsmuster besser abzubilden. Eine Überdeckung des Personalbedarfs in einzelnen Perioden, die sich im Ergebnis der Personaleinsatzplanung zeigt, ist unvermeidbar. Somit ist es insbesondere für kleine Call Center sinnvoll, die Überdeckung des Bedarfsmusters zu antizipieren. Sollte diese bei der Berechnung der Kennzahl keine Berücksichtigung finden, ist absehbar, dass sich bei der wöchentlichen Einsatzplanung der Personalbedarf nicht immer decken lässt. Wenn der Personalbedarf der einzelnen Perioden im Rahmen der Personaleinsatzplanung mindestens gedeckt werden soll, dann ist es notwendig, die aggregierten Personalbedarfsstunden des betrachteten Zeitraums um einen Puffer zu erhöhen, der die Überdeckung des Personalbedarfs als Ergebnis der Personaleinsatzplanung antizipiert. Dieser Puffer nimmt indirekt das Personalbedarfsmuster vorweg, das im Rahmen der operativen Personaleinsatzplanung zu decken ist.

Die Höhe des Zuschlags hat einerseits Implikationen auf die Personalkosten. Andererseits vermittelt er die Sicherheit, dass die Schichten der Vollzeitmitarbeiter das Bedarfsmuster abdecken können. Der Pufferzuschlag lässt sich anhand von Vergangenheitsdaten ermitteln, sofern diese vorliegen. Das bedeutet, dass man aus vergangenen Personaleinsatzplanungsproblemen eine durchschnittliche oder eine saisonale Abweichung zwischen Personalbedarf und kostenoptimalem Personaleinsatz berechnet. Diese Abweichung kann in Prozent ausgedrückt den Pufferzuschlag ergeben. Dabei ist jedoch zu beachten, dass sich bei diesem Vorgehen eine schlechte Planung der Vergangenheit, die sich in überhöhten oder aber auch zu geringen Pufferzuschlägen ausdrückt, in der zukünftigen Planung fortschreibt und weiterhin zu schlechten Ergebnissen führt.

Die Methode lässt insbesondere keinen Schluss auf die Anzahl der beschäftigten Agenten zu, da sich die Auswirkungen einer auf dieser Basis getroffenen Entscheidung nicht vollständig abschätzen lassen. Bei einem zugrunde gelegten Bezugszeitraum von einer Woche bzw. einem Monat werden aufgrund des unterschiedlichen Nachfrageniveaus eine verschiedene Anzahl von Vollzeitmitarbeitern für die einzelnen Wochen respektive Monate ermittelt. Inwieweit trotz der variierenden Ergebnisse der einzelnen Wochen bzw. Monate bezüglich der benötigten Vollzeitmitarbeiter die Anzahl der einzustellenden Agenten ermittelt werden kann bleibt offen. Bei einem Bezugszeitraum von einem Jahr wird eine Anzahl an Vollzeitäquivalenten der Variation des Nachfrageniveaus der Woche nicht gerecht. Denn implizit würde ein solcher Ansatz von der mittleren Anzahl der Vollzeitmitarbeiter im gesamten Planungszeitraum ausgehen.

Die Methode der Vollzeitäquivalente liefert zwar eine Aussage über den mittelfristigen Personalbedarf, sie ermöglicht jedoch aufgrund des schwankenden Personalbedarfs innerhalb eines Jahres keinen Schluss auf die Anzahl an beschäftigten Agenten. Insbesondere unterbleibt die Entscheidung über die Aufteilung der Verträge bezüglich Vollzeit- und Teilzeitagenten und deren dazugehörigen vertraglichen Arbeitsstunden. Da sich die Arbeitszeit eines Vollzeitagenten nicht linear in die mehrerer Teilzeitagenten aufteilen lässt, ist es schwierig, von den ermittelten Vollzeitäquivalenten auf eine Aufteilung in Vollzeit- und Teilzeitagenten zu schließen. Demnach ist der Pufferzuschlag zu korrigieren, sobald nicht ausschließlich Vollzeitagenten eingesetzt werden. Bekanntermaßen kann bei der Beschäftigung eines hohen Anteils an Teilzeitagenten das bestehende Personalbedarfsmuster mit einer geringeren Überdeckung befriedigt werden, so dass sich der Pufferzuschlag reduzieren lässt. Die Bestimmung der Höhe des Pufferzuschlages ist demnach

nicht ohne die Berücksichtigung des Anteils der Teilzeitarbeitnehmer möglich. Eine vorherige Festlegung des Anteils kann jedoch zu höheren Kosten führen als dies bei einer alternativen Aufteilung der Anteile von Vollzeit- und Teilzeitagenten der Fall ist. Vom Ergebnis des mittelfristigen Personalbedarfs ist demnach kein direkter Rückschluss auf den Personalbestand möglich.

Es wurde bereits erörtert, dass die Personalbestandsplanung der mittelfristigen Personalbedarfsbestimmung im Call Center vorzuziehen ist. Ein Call Center, das einen konstanten Personalbestand verfolgt, kann die Bestimmung der Struktur der Arbeitsverträge nicht losgelöst von den im Abschnitt 2.5.3.2 beschriebenen Anpassungsmaßnahmen durchführen. Nur eine gleichzeitige Betrachtung der Anpassungsmaßnahmen in Verbindung mit dem kurzfristigen Personalbedarf und der Bestimmung der Aufteilung an Vollzeit- und Teilzeitagenten stellt sicher, dass einerseits ein kostenminimaler Personalbestand im Call Center vorhanden ist. Andererseits garantiert er eine zielgerechte Personaleinsatzplanung auf der operativen Planungsebene. Somit stellt lediglich eine ganzheitliche Betrachtung der Anpassungsmaßnahmen sicher, dass die Arbeitszeiten in Antizipation der Schwankungen im Anrufvolumen variabel und kostenoptimal gestaltet werden können.

3.6 Zusammenfassung der Ergebnisse der Personalbedarfsermittlung

In diesem Kapitel wurden zunächst die Determinanten des Personalbedarfs beschrieben. Die Nachfrage beeinflusst primär den Personalbedarf. Eine gute Prognose der Nachfrage stellt die Grundvoraussetzung für die Personalbedarfsermittlung dar. Aus diesem Grunde erfolgte eine kurze Darstellung der für Call Center geeigneten Methoden zur Vorhersage der Nachfrage. Nach der Beschreibung der Anforderungen, denen die Methoden der Personalbedarfsermittlung genügen müssen, wurden die Methoden zur Ermittlung des kurz- und mittelfristigen Personalbedarfs vorgestellt.

Bei der kurzfristigen Personalbedarfsermittlung in Call Centern kommen traditionell Warteschlangenmodelle zum Einsatz. Aufgrund des schwankenden Anrufvolumens im Zeitablauf eines Tages ist es notwendig, die Personalbedarfsermittlung beim Einsatz von Warteschlangenmodellen für kurze Intervalle durchzuführen, für die angenommen werden kann, dass die Ankunftsrate annähernd konstant ist. Das Verhältnis zwischen dem Anrufvolumen und der Anzahl der eingesetzten Agenten gestaltet sich nicht linear, so dass für längere Intervalle, in denen die Ankunftsrate erheblichen Schwankungen unterliegt, nicht unmittelbar ein Warteschlangenmodell zur Bedarfsermittlung eingesetzt werden kann. Sofern die Ankunftsrate mehrerer Intervalle zusammengefasst wird, um anhand eines

Warteschlangenmodells einen einheitlichen Personalbedarf für diese Intervalle zu bestimmen, wird der tatsächlich benötigte Personalbedarf in einigen Perioden erheblich überschätzt, in anderen hingegen unterschätzt. Demnach lassen sich die Intervalle bei der Personalbedarfsermittlung als solche nicht aggregieren, um auf den mittelfristigen Personalbedarf zu schließen. Die aus dem Planungsschritt der kurzfristigen Personalbedarfsplanung resultierenden Ergebnisse hingegen sind aggregierbar, da bei ihnen bereits das nichtlineare Verhalten zwischen der Anzahl der Anrufe in einem Intervall und dem Personalbedarf Berücksichtigung gefunden hat. Die Aggregation der Ergebnisse des kurzfristigen Personalbedarfs können bei der mittelfristigen Personalbedarfsermittlung ausgenutzt werden. Dabei haben im Rahmen der mittelfristigen Personalbedarfsermittlung jedoch alle beschäftigten Personen die gleichen vereinbarten Arbeitszeiten innerhalb eines Planungszeitraums. Neben der Aussage über den mittelfristigen Personalbedarf lässt dieses Ergebnis demnach keine Schlussfolgerung auf einen Personalbestand zu, bei dem sich die Agenten hinsichtlich der vereinbarten Arbeitszeit differenzieren lassen und der auf die Nachfragestruktur des Call Centers zugeschnittenen ist.

Im folgenden Kapitel vier wird zunächst ein Überblick über die bisherigen Veröffentlichungen der Personalbestands- und Personaleinsatzplanung gegeben. Im Anschluss daran widmet sich das Kapitel fünf einem mathematischen Modell zur Personalbestands- und aggregierten Personaleinsatzplanung, das die oben genannten Aspekte miteinander verbindet.