Optimale Echtzeit-Personaleinsatzplanung für Inbound Call Center durch Approximation von Warteschlangenkennzahlen mit Künstlichen Neuronalen Netzen

Der Wirtschaftswissenschaftlichen Fakultät der

Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften

– Doctor rerum politicarum –

vorgelegte Dissertation

von

Dipl.-Math. Frank Köller

[redacted]

2007

Inhaltsverzeichnis

1 Einleitung **1**
1.1 Motivation . . . 1
1.2 Zielsetzung . . . 4
1.3 Vorgehensweise . . . 5

2 Warteschlangentheorie und Simulation **13**
2.1 Warteschlangensysteme . . . 15
2.1.1 Ankunftsprozess . . . 17
2.1.2 Warteraum und Warteschlange . . . 18
2.1.3 Bedienstrategie und Bedienprozess . . . 18
2.1.4 Warteschlangenkennzahlen . . . 19
2.1.5 Klassifizierung von Warteschlangensystemen . . . 21
2.1.6 Stationärer Zustand und Markovprozesse . . . 22
2.1.7 Systeme mit exponential-verteilten Zwischenankunftszeiten und Bedienzeiten . . . 24
2.2 Simulation von Warteschlangenmodellen . . . 29
2.2.1 Simulationsprozess . . . 32
2.2.2 Genauigkeit stochastischer Simulationen . . . 34
2.2.3 Beispielsimulation mit Arena . . . 34

3 Kontaktcenter **41**
3.1 Bedeutung von Kontaktcentern . . . 43
3.2 Komponenten von Kontaktcentern . . . 45
3.3 Call Center Marktentwicklung . . . 48

4 Personaleinsatzplanung und Schichtplanung in Kontaktcentern **53**
4.1 Personalbedarfsermittlung und -einsatzplanung . . . 54
4.2 Relevante Publikationen und Modelle . . . 57

5 Künstliche neuronale Netze **61**
5.1 Funktionsapproximation und überwachtes Lernen . . . 64
5.2 Approximationsfunktionen . . . 66
5.2.1 Dreilagige Perzeptrons ohne Direktverbindungen . . . 66
5.2.2 Dreilagige Perzeptrons mit Direktverbindung . . . 70
5.2.3 Vierlagige Perzeptrons ohne Direktverbindungen . . . 73
5.2.4 Vierlagige Perzeptrons mit Direktverbindungen . . . 77
5.3 Bewertung der Güte einer Approximationsfunktion . . . 80

5.4 Maximalkrümmung und Gesamtkrümmung 81

6 Reengineering des FAUN Neurosimulators 83
6.1 Historie und Vorarbeiten an der TU Clausthal 85
6.2 FAUN-Kernel 89
6.2.1 Dynamische Speicherallokation 89
6.2.2 Windows Portierung 90
6.2.3 CPU-Rechenzeit 90
6.2.4 Ausreißererkennung 91
6.3 FAUN-Oberfläche 93
6.4 Identifizierung leistungsfähiger Fortran-Compiler und BLAS-Bibliotheken 100
6.5 FAUN Web Frontend und die MPI-Version 104
6.6 FAUN und Grid Computing 106
6.6.1 Vorteile und Mehrwert des Grid Computing 107
6.6.2 FAUN Grid Computing Client 108
6.7 Krümmungstensoren 109
6.7.1 Krümmung dreilagiger Perzeptrons ohne Direktverbindungen . . . 110
6.7.2 Krümmung dreilagiger Perzeptrons mit Direktverbindungen . . . 112
6.7.3 Krümmung vierlagiger Perzeptrons ohne Direktverbindungen . . . 113
6.7.4 Krümmung vierlagiger Perzeptrons mit Direktverbindungen . . . 115
6.7.5 Implementierung der Krümmungstensoren in FAUN 117
6.7.6 Online Auswertung der Gesamtkrümmung mit Gnuplot 119
6.7.7 Empirische Leistungsbetrachtung 122
6.7.8 Analytische Leistungsbetrachtung 127
6.8 Der FAUN-Neurosimulator nach dem Reengineering 129

7 Personaleinsatzplanung mit Künstlichen Neuronalen Netzen 133
7.1 Simulationsprogramm 136
7.1.1 Anforderungen 137
7.1.2 Modellierung 138
7.1.3 Verteilungen und Zufallszahlen in Maple 140
7.1.4 Realisation 143
7.1.5 Programmtest 146
7.2 M/M/1-Modell und künstliche neuronale Netze 147
7.2.1 Simulation des M/M/1-Modells und die analytische Lösung 147
7.2.2 Training der künstlichen neuronalen Netze 153
7.2.3 Auswertung und Ergebnisse für das M/M/1-Modell 168
7.3 M/M/c-Modell und künstliche neuronale Netze 171
7.3.1 Simulation des Inbound Call Centers anhand des M/M/c-Modells 172
7.3.2 Genauigkeit der stochastischen Simulationen für das M/M/c-Modell 175
7.3.3 Approximation der M/M/c-Kennzahlen 177
7.3.4 Auswertung für das M/M/c-modellierte Inbound Call Center . . . 182
7.4 Arena-Simulation und künstliche neuronale Netze 184
7.4.1 Modellierung des Call Centers mit Arena 185

7.4.2 Durchführung der Simulationen mit Arena 186
7.4.3 Approximation der Kenngrößen des Call Centers 191
7.4.4 Auswertung des mit Arena simulierten Call Centers 194
7.5 Anwendung in einem realen Call Center 197
7.5.1 Datenaufbereitung und Datenanalyse 202
7.5.2 Simulation der Hotlines 210
7.5.3 Training der künstlichen neuronalen Netze 214
7.5.4 Auswertung in Maple 221
7.5.5 Praxiseinsatz der Neuronalen Netze 226
7.5.6 Reale Daten als direkte Grundlage für das Training 236

8 Kritische Würdigung 245
8.1 Kritische Würdigung zum Reengineering 245
8.2 Kritische Würdigung zur Personaleinsatzplanung 247

9 Zusammenfassung und Management-Summary 251
9.1 Zusammenfassung zum Reengineering 251
9.2 Management-Summary zur Personaleinsatzplanung 255

A Ableitungen der verschiedenen Perzeptrons 265
A.1 Ableitungen 3-lagige Perzeptron ohne Direktverbindungen 265
A.2 Ableitungen 3-lagige Perzeptron mit Direktverbindungen 267
A.3 Ableitungen 4-lagige Perzeptron ohne Direktverbindungen 269
A.4 Ableitungen 4-lagige Perzeptron mit Direktverbindungen 274

B Empirische Leistungsbetrachtung der Krümmung 279

C FAUN Quellcode der Krümmungsberechnung 287

D Maple Quellcode der Simulationsprogramme 293
D.1 Benutzeranleitung zum Simulationsprogramm für das M/M/1- Modell . . 293
D.2 Quellcode zum M/M/1- Modell 297
D.3 Quellcode zum M/M/c- Modell 301
D.4 Quellcode zum Call Center 304

E Abbildungen zu den Animationen 315

F Gesamtergebnisse der Überprüfung der neuronalen Netze im Echtzeitbetrieb 319

1 Einleitung

Gegenstand dieser Doktorarbeit ist die Personaleinsatzplanung in Kontaktcentern bzw. Kundenservice Centern. Die Personaleinsatzplanung hängt von vielen verschiedenen betriebswirtschaftlichen Faktoren[1] ab, die einen komplexen Zusammenhang bilden, vgl. [131]. Bei den bisher zur Personaleinsatzplanung eingesetzten Modellen werden diese komplexen Zusammenhänge und Strukturen vereinfacht, um überhaupt zu einer Lösung und zu einem optimierten Kontaktcenterbetrieb zu kommen, vgl. [91]. Das Verfahren birgt die Gefahr, dass Ergebnisse aufgrund der Vereinfachung ungenau werden. Der Autor beschäftigt sich bereits seit längerem mit der Anwendung künstlicher neuronaler Netze bei schwierigen, nichtlinearen, multivariaten Approximationsproblemen. In dieser Arbeit wird untersucht, ob die Personaleinsatzplanung als komplexes Thema mit Hilfe künstlicher neuronaler Netze optimiert werden kann, so dass im Vorfeld keine Vereinfachung der Einflussfaktoren vorgenommen werden muss und die Ergebnisse möglichst genau und realitätsnah sind.

1.1 Motivation

Kontaktcenter haben in Unternehmen eine besondere Schlüsselposition inne: Sie bilden die Schnittstelle zwischen Unternehmen und Kunden. Es werden sowohl im B2B- als auch im B2C-Bereich Geschäftsbeziehungen hergestellt, gesteuert und weiter ausgebaut. Kontaktcenter bieten, im Gegensatz zu reinen Call Centern, bei denen nur per Telefon mit den Kunden kommuniziert wird, so genannte Teleservices für entfernte Kunden an, z. B. über das Internet, per E-Mail, Fax, Telefon oder über andere Kommunikationskanäle[2]. Generell wird unterschieden, ob der Kunde aktiv auf das Unternehmen zugeht (Inbound), oder ob das Unternehmen den Kontakt zum Kunden aufgenommen hat (Outbound). Neben den Kontaktcentern, die alle Wege der Kommunikation unterstützen, gibt es auch reine Inbound und Outbound Call Center, die hauptsächlich auf den telefonischen Kontakt mit den Kunden ausgelegt sind.

Da heute in vielen Unternehmen bereits 90 % aller Kundenkontakte in Kontaktcentern abgewickelt werden und dieser Unternehmensbereich maßgeblich zur gesamtunternehmerischen Wertschöpfung beiträgt, ist es für den Gesamterfolg des Unternehmens unabdingbar, dass alle Abläufe im Kontaktcenter permanent gesteuert und optimiert werden, vgl. [43]. Hierbei sind aus Unternehmenssicht Kosten und Performance maßgebliche Erfolgsfaktoren.

[1] Hier werden vor allem Überlegungen zum Kosten-Nutzenverhältnis getätigter Maßnahmen angestellt.

[2] Andere Kommunikationskanäle können SMS oder MMS sein, aber auch zukünftig Teletext, wenn über das TV-Kabel ein Rückkanal zur Verfügung steht.

Die Kosten setzen sich aus fixen Kosten wie Miete und Ausstattung sowie variablen Kosten wie laufende Betriebskosten und Personalkosten[3] zusammen. Die Personalkosten bilden mit etwa drei Vierteln den überwiegenden Anteil des Gesamtbudgets in einem Kontaktcenter, vgl. [43, 82, 113]. Die verbleibenden Positionen weisen jeweils nur eine untergeordnete Größenordnung auf und können in der Praxis auch nicht weiter gesenkt werden, vgl. [105].

Für den Kunden ist es besonders wichtig, wie schnell auf seine Anfragen per Fax oder E-Mail geantwortet wird, bzw. wie lange er bei einem Telefonanruf in der Warteschleife verbringt, bis ein Kundenberater frei ist. Dies bedeutet, dass ein hohes Niveau von Kundenzufriedenheit durch eine hinreichend große Anzahl an Kundenberatern erreicht werden kann. Dadurch verringert sich die Wartezeit der Kunden enorm und erhöht die Kundenzufriedenheit, verursacht aber auch höhere Kosten, vgl. [21]. Zudem müssen die Kundenberater entsprechend geschult sein, damit die Beratungsqualität hoch ist und der Kunde sich gut beraten fühlt. Welchen Dienstleister der Kunde nutzt, hängt demnach vom Servicelevel ab. Nach dem Minimalprinzip wird das Unternehmen gehalten sein, den angestrebten Servicelevel unter Entstehung möglichst geringer Kosten zu gewährleisten.

Während Ende des letzten Jahrtausends noch viele Call Center entstanden sind, in denen maximaler Service ohne Rücksicht auf die Kosten geboten wurde, expandiert heute die einstige Boom-Branche nicht mehr, bedingt durch wirtschaftlichen Stillstand, Kostendruck und dennoch hohen Service-Erwartungen, vgl. [105]. Das steigende Kommunikationsaufkommen in den vergangenen Jahren und der unreflektierte maximale Service in Call Centern verursachten eine Kostenexplosion, die in keinem angemessenen Verhältnis zur Umsatzentwicklung steht, vgl. [43]. Somit stehen Call Center nun vor der konkreten Aufgabe, Maßnahmen zur Kostensenkung aktiv umzusetzen. Dazu muss das angebotene Servicespektrum an die tatsächlichen Bedürfnisse angepasst und gleichzeitig die Effizienz der Prozesse bzw. die Auslastung der Agenten[4] gesteigert werden. Daher wird eine Personalbedarfs- und Personaleinsatzplanung erstellt, welche für einen vorgegebenen Servicegrad hinsichtlich der Wartezeit die erforderliche Zahl von Kundenberatern je Periode bzw. Schicht angibt. Hierbei wird in Kauf genommen, dass unter Umständen nicht jeder Anrufer sofort bedient und in eine sogenannte Warteschlange geschaltet wird.

Bei der Fokussierung auf Einsparpotenziale, wie die Freisetzung der tatsächlich entbehrlichen Kapazitäten, darf das Call Center Management jedoch nicht die Kundenzufriedenheit gefährden. Im Rahmen von Kosten-Nutzen-Analysen muss ermittelt werden, bei welchem Personalaufwand die Wirtschaftlichkeit optimiert werden kann, so dass der Return on Investment steigt[5].

In der Praxis wird bei der Personaleinsatzplanung in Call Centern meist noch das M/M/c- (oder Erlang-C) Warteschlangenmodell unterstützend zu anderen Methoden[6]

[3] Die Kosten für das Personal setzen sich u. a. aus Arbeitsentgelt, Personalzusatzkosten, Personalauswahl sowie Schulungen und Training zusammen.

[4] Die Kundenberater eines Call Centers werden meist als Agenten bezeichnet.

[5] Dieser Aspekt wird im Weiteren nicht mehr behandelt und es wird nur auf die benötigte Anzahl an Agenten eingegangen.

[6] Vgl. z. B. das Erlang-A Modell als Erweiterung des Erlang-C Modells in [70, 127], welches realistischer für die Kontaktcentermodellierung ist. Weitere Methoden werden in [5, 92, 113, 125, 126, 137, 214]

verwendet. Da für das M/M/c-Modell analytische Lösungen für alle Warteschlangenkennzahlen[7] existieren, ist es einfach und schnell anwendbar. Demgegenüber stehen einschränkende Annahmen des M/M/c-Modells, z. B., dass alle Anrufer geduldig sind, der Warteraum unendlich groß ist und dass die Zwischenankunftszeiten ebenso wie die Bearbeitungszeiten unabhängig exponentialverteilt sind. Auf viele Inbound Call Center treffen diese Annahmen des Erlang-C Modells aber nicht zu, vgl. [5, 92, 125]:

- In der Regel steht nur eine begrenzte Zahl an Wartepositionen zur Verfügung, das heißt, wenn der Warteraum voll ist, erhält der Anrufer ein Besetztzeichen.

- Call Center weisen meist mehrere Klassen von Anrufern und Agenten auf.

- Einige der Anrufer sind in der Warteschleife ungeduldig und legen vorzeitig auf, um zu einem späteren Zeitpunkt eventuell wieder anzurufen. Unter Umständen erfolgen sogar mehrere Anrufversuche.

Sind die Zwischenankunfts- und die Bearbeitungszeiten nicht exponentialverteilt, so ist es nur schwer bzw. gar nicht möglich, eine analytische Lösung zu finden. Dies gilt insbesondere für Kontaktcenter, wo zusätzlich noch andere Kommunikationswege als das Telefon benutzt werden.

Das M/M/c-Modell gibt die Realität nur ungenau wieder, vgl. [214]. Dennoch können grundlegende Zusammenhänge auf der Basis dieses einfachsten Modells in konzeptionell klarer Weise erläutert werden und daher wird es regelmäßig bei der Personaleinsatzplanung in der Praxis eingesetzt. Kann keine analytische Lösung für ein Warteschlangenmodell gefunden werden[8], so werden Näherungen für die Warteschlangenkennzahlen aufgrund von aufwändigen Simulationen, z. B. mit der Simulationssoftware Arena, berechnet.

Die Berechnung großer, umfangreicher Simulationsmodelle, die von vielen stochastischen Variablen abhängen, ist nur schlecht möglich. Der Zeitaufwand für die Generierung hinreichend genauer, brauchbarer Warteschlangenkennzahlen allein durch die Simulation wäre enorm. Dementsprechend können nur kleinere Modellinstanzen des gesamten Problems einzeln mit Simulationen in einem einigermaßen wirtschaftlichen Rahmen untersucht werden. Die Ergebnisse müssen dann noch zusammengefügt und neu interpretiert werden. Deshalb wird in dieser Arbeit eine neue Möglichkeit vorgestellt, wie auch umfangreichere Modelle in adäquater Zeit gelöst werden können.

---

vorgestellt.

[7] Wichtige Warteschlangenkennzahlen für Call Center Manager sind z. B. die durchschnittliche Wartezeit der Kunden und der Auslastungsgrad der Call Center Agenten.

[8] Diese Warteschlangenprobleme, bei denen keine exakten, explizit analytischen Lösungen angegebenen werden können, gelten als schwierige Warteschlangenprobleme. Gleiches gilt für Systeme, die zu komplex sind und stark vereinfacht werden müssten, damit eine analytische Lösung berechnet werden kann, vgl. [11, 22, 77].

1.2 Zielsetzung

Diese Ausführungen verdeutlichen, dass die Messung der Performance und die Bestimmung der relevanten Kennzahlen in Kontaktcentern ein bekanntes, aber noch nicht zufriedenstellend gelöstes Problem ist. Bereits eine geringe Abweichung zwischen den eingesetzten und den tatsächlich benötigten Agenten führt zu immensen Kosten. Hieraus ergibt sich die Frage, wie eine bessere Personaleinsatz- und Schichtenplanung zu den bisherigen, unzureichend genauen Methoden erreicht werden kann.

Künstliche neuronale Netze sind in der Lage, hochdimensionale, komplexe Input- und Outputzusammenhänge zu erlernen, ohne dass die Struktur schwieriger Problemstellungen vereinfacht werden muss, vgl. [33]. Das Problem der Personaleinsatz- und Schichtplanung bzw. die Bestimmung der relevanten Kennzahlen lässt sich mathematisch als nichtlineares multivariates Approximationsproblem formulieren. In der Praxis werden künstliche neuronale Netze bei nichtlinearen multivariaten diskreten Approximationsproblemen sehr erfolgreich eingesetzt und haben sich gegenüber anderen Verfahren durchgesetzt, vgl. [33]. Die Möglichkeit, Personaleinsatz- und Schichtenplanungen in Call Centern mit Hilfe künstlicher neuronaler Netze vorzunehmen, wird in dieser Arbeit zum ersten Mal untersucht. Insbesondere wird folgenden **Forschungsfragen** nachgegangen, welche auch gleichzeitig die Vorgehensweise in dieser Arbeit repräsentieren:

- Können mit künstlichen neuronalen Netzen Warteschlangenkennzahlen approximiert werden?
- Welche Ergebnisse erzielen künstliche neuronale Netze im Vergleich zu den analytischen Lösungen bei einfachen Warteschlangenproblemen?
- Können künstliche neuronale Netze auch die Warteschlangenkennzahlen bei nicht analytisch lösbaren Warteschlangenproblemen approximieren?
- Kann mit künstlichen neuronalen Netzen eine realistische Personaleinsatz- und Schichtenplanung im Echtzeitbetrieb vorgenommen werden?
- Können aus den so gewonnenen Ergebnissen Handlungsempfehlungen für Call Center Manager abgeleitet werden?

Um eine bessere Personaleinsatz- und Schichtenplanung in Kontaktcentern und Call Centern zu erstellen, wird also in dieser Arbeit der **zentralen Forschungsfrage** nachgegangen, ob und wie gut künstliche neuronale Netze in der Lage sind, Warteschlangenkennzahlen zu approximieren. Dieser Zusammenschluss der Forschungsfelder der künstlichen Intelligenz und der Warteschlangentheorie wurde bisher in der Forschung und Literatur nicht behandelt[9]. Die Arbeit leistet damit einen Beitrag zur Schließung dieser Forschungslücke.

[9] In [100] werden neuronale Netze lediglich zur Prognose des Anruferaufkommens benutzt, nicht aber zur genauen Bestimmung der Warteschlangenkennzahlen.

1.3 Vorgehensweise

Nach einer Einführung in die Warteschlangentheorie und Simulationstechniken in Kapitel 2 werden verschiedene Kontaktcenterausprägungen in Kapitel 3 und deren Einfluss auf die unternehmerischen Ziele vorgestellt. In Kapitel 4 wird auf die Personaleinsatzplanung und Schichtplanung in Kontaktcentern eingegangen. Danach erfolgt die Vorstellung der künstlichen neuronalen Netze und des verwendeten Neurosimulators in den Kapiteln 5 und 6.

Die entsprechenden Grundlagen für schwierige Warteschlangenprobleme, bei denen keine analytischen Lösungen existieren, werden dadurch geschaffen, dass zunächst Warteschlangenmodelle, wie das M/M/1- oder das M/M/c-Modell, bei denen die analytischen Lösungen einfach berechnet werden können, mit künstlichen neuronalen Netzen untersucht werden. Die daraus resultierenden Ergebnisse können dann auf die schwierigen Warteschlangenprobleme übertragen werden, vgl. Abbildung 1.1 und Kapitel 7. Dabei muss, im Vergleich zum Lösungsansatz mit aufwändigen Simulation, z. B. nicht die grundlegende Struktur der Problemstellung verändert werden.

Anhand von Simulationen für Inbound Call Center wird im Rahmen des Vorhabens gezeigt, dass künstliche neuronale Netze Kennzahlen von Warteschlangenproblemen, bei denen analytische Lösungen existieren, sehr gut approximieren können. Dazu werden mit Hilfe der Simulationssoftware Arena und der Computeralgebrasoftware Maple Simulationen für die verschiedensten Kennzahlen gefahren, um Muster für das Training der künstlichen neuronalen Netze zu erzeugen. Der zusätzliche Schritt des Trainings, der nur wenige Sekunden dauert, erfolgt dann mit dem Neurosimulator FAUN (Fast Approximation with Universal Neural Networks), der derzeit am Institut für Wirtschaftsinformatik der Leibniz Universität Hannover weiterentwickelt wird[10]. Im Vergleich zu einem Lösungsansatz, der nur auf der reinen Simulation beruht, ist das Training der künstlichen neuronalen Netze nach der Mustergenerierung durch eine Simulation ein weiterer Schritt, der sich dadurch rechtfertigt, dass mit einem geringeren zusätzlichen Aufwand eine erhebliche Ergebnisverbesserung erreicht werden kann. Diese Untersuchung der analytisch lösbaren Systeme ist wichtig, da so die Abweichungen der Ergebnisse der künstlichen neuronalen Netze zu den analytischen Lösungen der Warteschlangenkennzahlen bestimmt werden können, vgl. auch Abbildung 1.2. Zusätzlich bilden weitere Erkenntnisse bei der stochastischen Analyse, wie z. B. welche Netztopologie die geeignetste ist, oder dass nur wenige, verrauschte Simulationspunkte ausreichen, um sehr gute künstliche neuronale Netze zu trainieren, die die Problemstellung hinreichend genau erlernt haben[11], die gesicherte Grundlage dafür, dass in einem weiteren Schritt künstliche neuronale Netze auch auf allgemeine Warteschlangenprobleme angewendet werden können, für die keine exakten, expliziten Lösungen für die Warteschlangenkennzahlen existieren. Meist können bei diesen Problemen ohne exakte Lösungen obere und untere Schranken bestimmt werden,

[10] Der Autor ist der Projektmanager der FAUN-Projektgruppe des Instituts für Wirtschaftsinformatik und maßgeblich an der Weiterentwicklung und des Reengineerings des FAUN Neurosimulators beteiligt, siehe dazu auch das entsprechende Kapitel 6.

[11] Hinreichend genau bedeutet hier, dass das neuronale Netz fast deckungsgleich mit der analytischen Lösung ist.

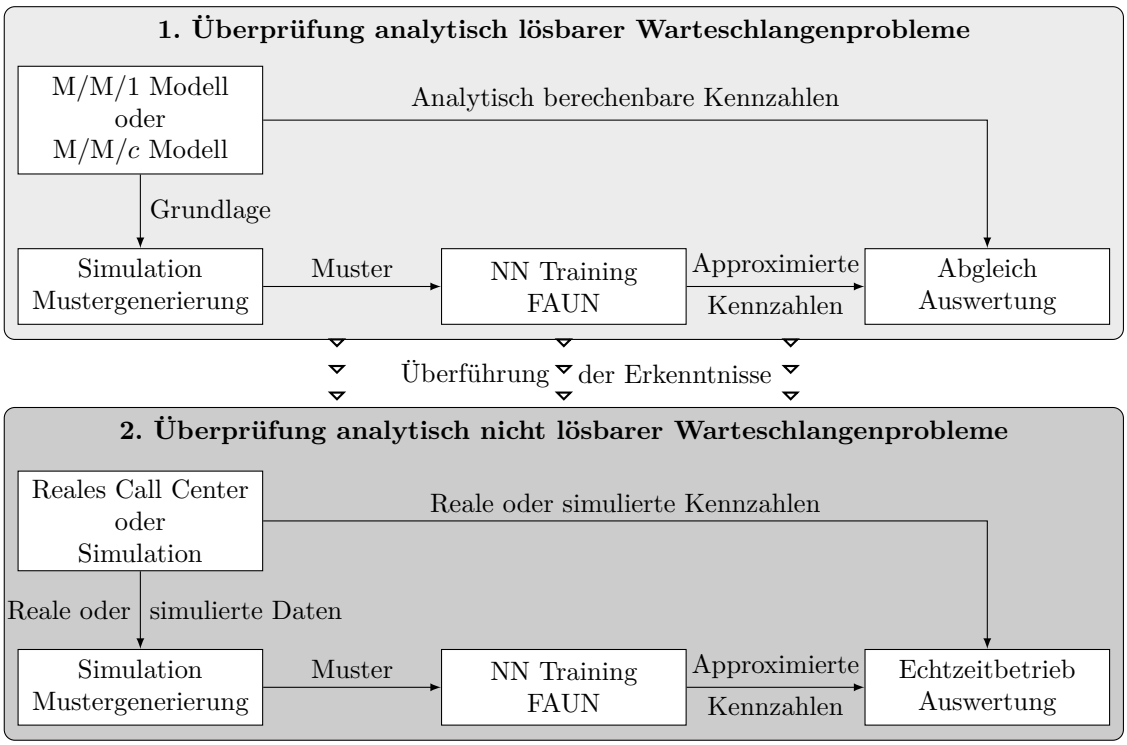

Abbildung 1.1: Um zu überprüfen, ob mit neuronalen Netzen (NN) Kennzahlen von Warteschlangenproblemen approximiert werden können, werden zunächst analytisch lösbare Warteschlangenprobleme untersucht. In einem weiteren Schritt werden dann die Erkenntnisse auf schwierige Warteschlangenprobleme überführt und deren Warteschlangenkennzahlen bestimmt. Am Beispiel eines simulierten oder realen Call Centers ist es möglich, die approximierten Kennzahlen mit den simulierten bzw. realen Daten im Echtzeitbetrieb abzugleichen.

die die Bandbreiten für Warteschlangenkennzahlen eingrenzen. Somit ist überprüfbar, ob die approximierten Kennzahlen innerhalb dieser Bandbreiten liegen [146]. Ein weiterer wichtiger Punkt ist, dass diese schwierigen Warteschlangenprobleme oft nur mit aufwändigen Simulationen gelöst werden können. Die Lösung besteht dann aus vielen einzelnen Punkten. Im Gegensatz zu einer flächendeckenden Auswertung werden bei der Anwendung von künstlichen neuronalen Netzen nur wenige Simulationspunkte für das Training benötigt. Diese Simulationspunkte können, bedingt durch die Glättungseigenschaften von künstlichen neuronalen Netzen, durchaus verrauscht sein. Zudem sind die approximierten Warteschlangenkennzahlen stetige, kontinuierlich auswertbare Funktionen, die auch zwischen den Simulationspunkten sinnvolle Ergebnisse liefern.

Die Untersuchung schwieriger Warteschlangenprobleme ohne analytische Lösungen für die Warteschlangenkennzahlen erfolgt ähnlich zu den Systemen mit analytischen Lösungen. Der wesentliche Unterschied besteht jedoch darin, dass reale Daten die Basis zur Mustergenerierung für das Training der künstlichen neuronalen Netze bilden. Es werden

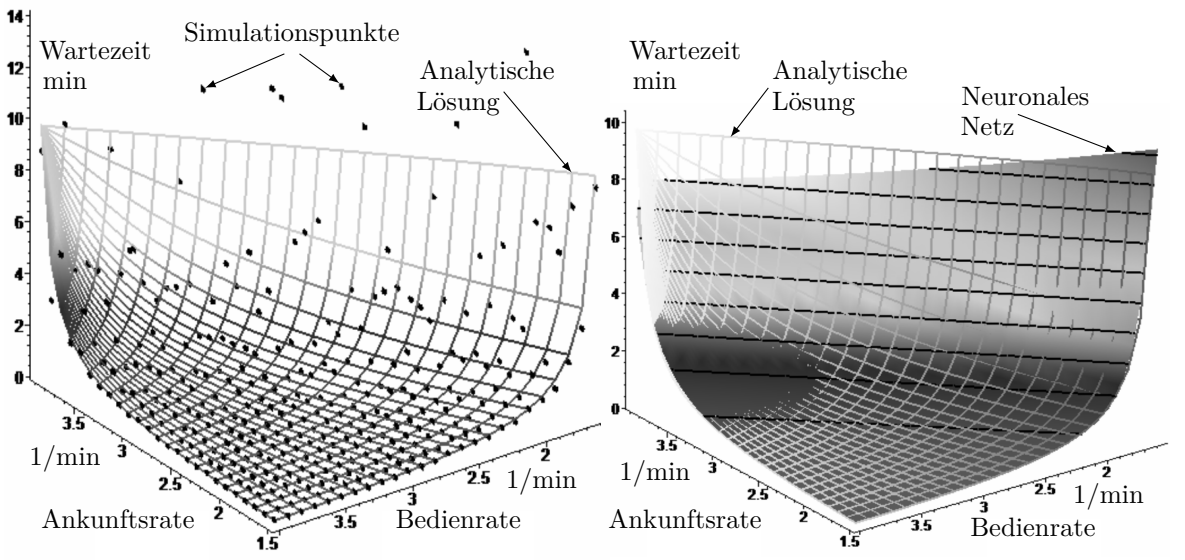

Abbildung 1.2: Links sind die Punkte der Simulation, die zum Training der durchschnittlichen Wartezeit der Kunden im System für die neuronalen Netze benötigt werden, und die tatsächliche analytische Lösung (Gitternetz) des M/M/c-Warteschlangenmodells dargestellt. Es ist so möglich, die Genauigkeit der Simulation zu erfassen. Rechts ist die analytische Lösung (Gitternetz) zu dem neuronalen Netz (Fläche mit Höhenlinien) zu sehen. Das neuronale Netz ist eine sehr gute Annäherung an die analytische Lösung für die Wartezeit der Kunden im System und hat damit die Problemstellung hinreichend genau erlernt.

verschiedenste Hotlines des Call Centers BHW Direktservice GmbH[12] ausgewertet, um praxisnahe Verteilungsfunktionen für den Ankunfts- und Bedienprozess abzuleiten. Diese Verteilungsfunktionen bilden die Grundlage für die Simulation realistischer Muster, die wiederum für das Training der künstlichen neuronalen Netze für die verschiedenen Warteschlangenkennzahlen verwendet werden, vgl. Abbildung 1.3. Der zusätzliche Schritt der Simulation ist wichtig, da so auch Systeme mit von der Realität abweichenden Zuständen, z. B. Erhöhung oder Verringerung der Anzahl an Agenten, untersucht werden können. In dem betrachteten Call Center werden immer die gleichen Anzahlen an Agenten zu bestimmten Ankunftsraten der Anrufer und Uhrzeiten eingesetzt. Die Simulation abweichender Agentenanzahlen kann zu Verbesserungen der Performance des Call Centers und zu Kosteneinsparungen führen. Natürlich können die realen Daten direkt als Basis für das Training der künstlichen neuronalen Netze dienen, jedoch könnte dann nur begrenzt eine Aussage über die Auswirkungen beim Ausbau oder Abbau der Kapazitäten im Call Center gemacht werden. Zudem kann durch die Simulation auch unnatürlich hohes oder niedriges Anrufaufkommen zu Zeiten, wo eigentlich nur wenig

[12]Eine Kooperation mit der BHW Direktservice GmbH mit Standort Hameln, dem Call Center der BHW Gruppe, die von der Deutschen Postbank AG übernommen wurde.

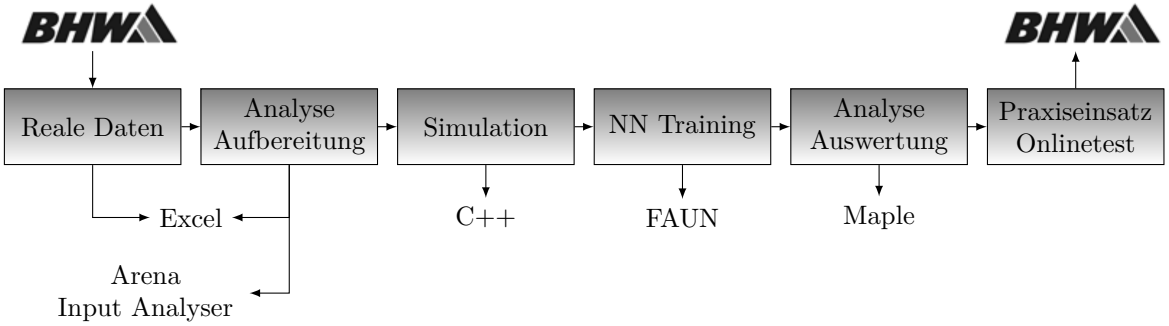

Abbildung 1.3: Um einen unterstützenden Echtzeitbetrieb in einem Call Center wie dem der BHW Gruppe zu realisieren, werden reale Daten in Excel und dem Input Analyser von der Simulationssoftware Arena ausgewertet. Die so gewonnenen Verteilungen für die verschiedenen Prozesse in dem Call Center werden dann einer Simulation, die die Muster für das Training der neuronalen Netze (NN) erzeugt, zur Verfügung gestellt. Die neuronalen Netze können nach der Analyse mit der Computeralgebrasoftware Maple im Praxiseinsatz unter realen Bedingungen getestet werden.

oder viele Anrufe eingehen, besser berücksichtigt und so mittrainiert werden. Das so bestimmte Modell wird dadurch flexibler und robuster gegenüber Störungen.

In einem weiteren Schritt wird dann nach der Bestimmung der verschiedenen Warteschlangenkennzahlen für die einzelnen Hotlines[13] die benötigte Anzahl an Agenten des First Levels, also der Bereich des Call Centers, in dem zunächst alle Gespräche dieser Hotlines auflaufen, zu einem vorgegebenen Servicelevel bestimmt. Dieser beträgt bei der BHW Direktservice GmbH 70/20, d. h., dass 70 % der eingehenden Anrufe innerhalb der ersten 20 Sekunden von den Agenten angenommen werden müssen. Bei der Bestimmung der Agentenanzahl ist nicht nur der zu erzielende Servicelevel ein entscheidender Inputparameter für die künstlichen neuronalen Netze, sondern auch die Uhrzeit und die dementsprechende Ankunftsrate der eingehenden Anrufe. Die Ankunftsraten in Call Centern können über den Tag gesehen stark variieren und sind dementsprechend abhängig von der Uhrzeit, dem Wochentag und sogar vom Monat. Zudem können Marketingkampagnen, wie z. B. aktuell ausgestrahlte Fernsehwerbespots oder Werbeschreiben, die Ankunftsraten zusätzlich erhöhen und so das System unnatürlich stören, vgl. [91]. Dies führt zu teils schwer vorhersehbaren Spitzen bei den Ankunftsraten der Anrufer bestimmter Hotlines, je nachdem, wie die Kunden auf die entsprechenden Marketingstrategien reagieren. Diese zusätzlichen Einflüsse müssen von den künstlichen neuronalen Netzen beim Echtzeitbetrieb abgefangen werden und setzen eine gewisse Robustheit der Netze gegenüber Störungen voraus. In mehreren Tests und unterstützendem Echtzeitbetrieb vor Ort im BHW Call Center sind die künstlichen neuronalen Netze zur Bestimmung der Agentenanzahl unter realen Bedingungen mit den tatsächlich vorhandenen Agenten-

[13] In dem betrachteten Call Center stehen dem Kunden unterschiedliche Hotlines zur Verfügung, die auch unterschiedlich frequentiert sind.

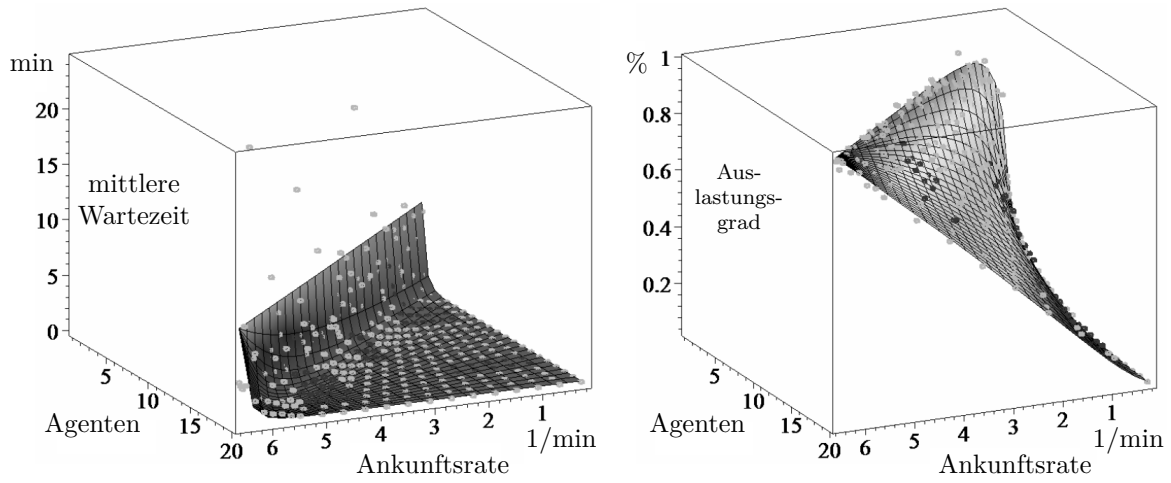

Abbildung 1.4: Neben der durchschnittlichen Wartezeit der Kunden ist der Auslastungsgrad der Agenten für einen Call Center Manager von entscheidender Bedeutung. Links ist die approximierte durchschnittliche Wartezeit der Kunden in der Warteschleife und rechts der dazugehörige approximierte Auslastungsgrad mit den jeweiligen Simulationspunkten zu sehen.

anzahlen abgeglichen worden. So wurden eine Machbarkeitsstudie und eine analytische Auswertung erstellt.

Neben dem Servicelevel und der durchschnittlichen Wartezeit der Kunden ist für einen Call Center Manager die Auslastung seiner Agenten von entscheidender Bedeutung, vgl. Abbildung 1.4. Es wird versucht, einen möglichst hohen Auslastungsgrad der Agenten zu erreichen, denn eine geringe Auslastung der Agenten führt zu unnötigen Personalkosten. Insgesamt ist jedoch zu berücksichtigen, dass die Agenten nicht überlastet werden. Bei einem Auslastungsgrad von nahezu 100 % entstehen ohne Pausen gesundheitliche Risiken für die Agenten und die Ausfallrate wäre entsprechend hoch, vgl. [92]. Dies wiederum hat negative monetäre Auswirkungen, welches sich auch durch Theorien der Personaleinsatzplanung zur optimalen Auslastung der Agenten belegen lässt, vgl. [91].

Mit Hilfe von künstlichen neuronalen Netzen können diese Warteschlangenkennzahlen, wie die durchschnittliche Wartezeit und die Auslastung der Agenten, zur Entscheidungsfindung bei der Personaleinsatzplanung Call Center Managern zur Verfügung gestellt werden. Der Service kann so, trotz eventueller Personaleinsparungen, optimiert werden, ohne dass die Agenten überlastet werden.

Für die meisten in der Praxis vorkommenden Warteschlangenprobleme existieren keine exakten, expliziten Lösungen für die Warteschlangenkennzahlen. Diese werden entweder mit aufwändigen, diskreten Simulationen gelöst oder aber das Grundproblem wird soweit vereinfacht, dass es analytisch lösbar wird. Im Gegensatz dazu ist der Vorteil beim Training künstlicher neuronaler Netze, dass die Struktur des Problems nicht verändert werden muss. Weiterhin brauchen auch nur wenige Simulationspunkte gegenüber einer „flächendeckenden“ Auswertung mit einer Simulation generiert werden, da das unvermeidliche Rauschen in den Simulationsdaten (vgl. Kapitel 2.2.2) durch die konti-

nuierliche, approximierte Lösung geglättet wird, d. h. die Kennzahlen genauer verfügbar sind. Aufgrund deutlich weniger Simulationen besteht ein erheblicher Zeitvorteil, denn der zusätzliche Schritt des Trainings der neuronalen Netze dauert i. d. R. nur wenige Sekunden.

Da verschiedenste Warteschlangenmodelle mit künstlichen neuronalen Netzen anhand eines Inbound Call Centers untersucht werden, können die Erkenntnisse auch auf andere schwierige Problemstellungen überführt werden; z. B. aus der Produktion oder Logistikketten, aber auch auf andere Servicestationen, wo Kunden warten müssen, wie z. B. am Bank- oder Postschalter. Diese Doktorarbeit richtet sich dementsprechend nicht nur an Call Center Manager, sondern auch an alle, die komplexe Warteschlangenprobleme einfach und schnell lösen müssen, um z. B. ihre Wertschöpfungsketten in Produktion oder Logistik zu optimieren.

Ein Reengineering der Software war wichtig, um

- den FAUN Neurosimulator überhaupt für die in der Arbeit vorgestellte Problemklasse einzusetzen;
- komfortabel und effizient mit dem FAUN Neurosimulator zu arbeiten.

Die Entwicklung des Neurosimulators FAUN begann 1997 an der TU Clausthal und wird mit der FAUN-Projektgruppe durch den Autor am Institut für Wirtschaftsinformatik der Leibniz Universität Hannover weitergeführt. Heutige Neurosimulationen basieren auf kompletten Software Emulationen. In der Software sind alle entscheidenden Variablen eingebettet, wie zum Beispiel die Eingabe- und Ausgabeneuronen sowie weitere, so genannte innere Neuronen, die Verbindungen zwischen den Neuronen, so genannte Synapsen, und deren Gewichte. Es gibt viele verschiedene kommerzielle, aber auch lizenzfreie (Free- und Shareware) Neurosimulatoren, von denen sich der FAUN Neurosimulator durch spezielle Funktionalität abhebt.

Um diese spezielle Funktionalität zu erreichen, war eine vorbereitende Tätigkeit die völlige Überarbeitung des FAUN Neurosimulators Kernels. Bei den ersten, auf Fortran 77 basierenden FAUN Versionen, die ausschließlich auf Unix und Linux Rechnern liefen, musste noch der gesamte Quellcode durch ein Preprozessorprogramm der Problemstellung angepasst und bei jeder Änderung neu kompiliert werden. Das heißt, dass der Benutzer zusätzlich einen Compiler benötigte. Dies ist dadurch bedingt, dass unter Fortran 77 keine dynamische Speicherallokation wie unter Fortran 90 und 95 möglich ist. Folglich war die Übersetzung auf Fortran 95 Quellcode mit dynamischer Speicherallokation ein wesentlicher Schritt zur komfortableren und einfacheren Bedienbarkeit, die das Preprozessorprogramm überflüssig machten. Der neue Kernel passt sich automatisch dem Approximationsproblem an und belegt selbst dynamisch den Speicher.

Die Portierung auf das Betriebssystem Microsoft Windows und die Entwicklung einer einfach zu bedienenden Borland Delphi Oberfläche waren die nächsten Schritte. Die textbasierte Eingabe der Steuerparameter mit einem Editor kann so entfallen. Zudem wurde durch Identifikation leistungsstarker Compiler unter Linux und Windows Systemen eine gute Gesamtperformance und moderate Allokation der Ressourcen erreicht.

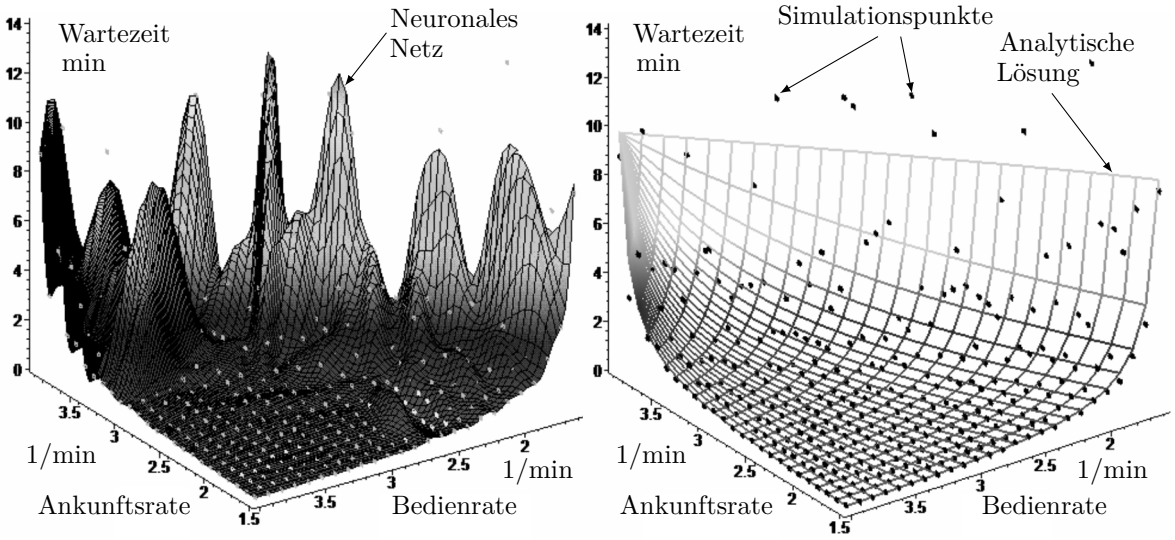

Abbildung 1.5: Das neuronale Netz (links) hat die Trainingsmuster nahezu auswendig gelernt und dadurch einen sehr geringen Trainingsfehler. Es oszilliert aber stark zwischen den Mustern und ist somit keine gute Annäherung an die hier ermittelbare analytische Lösung (rechts) für die Wartezeit der Kunden im System. Mit dem Gütemaß der Gesamtkrümmung können solche neuronalen Netze identifiziert werden.

Die neueste FAUN Version läuft auf Intel- und AMD-Maschinen bis zu 3-mal schneller als die ursprünglichen ersten Unix, Linux und Windows Versionen.

Wichtigster Bestandteil des FAUN Sofwarereengineerings war jedoch die Implementierung der Berechnung der Gesamtkrümmung künstlicher neuronaler Netze. Bisher waren die einzigen implementierten Gütemaße für ein approximiertes neuronales Netz die Trainings- und Validierungsfehler und deren Validierungsgüte. Auch, wenn eine vollautomatische Vermeidung des Übertrainierens der neuronalen Netze durch anpassbare Abbruchkriterien bereits im FAUN Neurosimulator implementiert ist, können immer noch neuronale Netze gefunden werden, die einen sehr niedrigen Trainings- und Validierungsfehler aufweisen, aber dennoch nicht die gewünschte Glattheit besitzen, vgl. Abbildung 1.5. Künstliche neuronale Netze, insbesondere mit mehreren verdeckten Neuronen, neigen dazu, zwischen den Mustern zu oszillieren. Diese Netze haben dann die Problemstellung nicht hinreichend genau erlernt, obwohl sie sehr niedrige Fehler aufweisen. Mit dem zusätzlichen Gütemaß der Gesamtkrümmung können nun solche oszillierenden Netze auch im hochdimensionalen Raum identifiziert und dementsprechend „glattere“ neuronale Netze mit einem geringfügig höheren Fehler bevorzugt werden, die aber die Problemstellung hinreichend genau erlernt haben, vgl. Abbildung 1.2 mit Abbildung 1.5.