

DAS MAGAZIN DER GOR

OR

News

NR.38 / MÄRZ 2010 / ISSN 1437-2045

Optimierung im Paketzentrum
Prognose mit neuronalen Netzen
Frutado – Advanced Planning Systems

Es braessiert!

GESELLSCHAFT FÜR OPERATIONS RESEARCH e.V.



GOR

GOR-Dissertationspreis 2009:

Simultane Bedarfsprognose und Warendisposition mit künstlichen Neuronalen Netzen

SVEN F. CRONE, LANCASTER

Der Planung wirtschaftlicher Beschaffungsmengen und Zeitpunkte fällt im Handel eine besondere Bedeutung zu: zu hohe Warenbestände führen zu einer Bereitstellung überflüssiger Güter, die als Überbestände Kapital binden, wertvollen Raum in Verkaufsflächen und Lagern belegen und aufgrund des Risikos von Beschädigungen, Diebstahl und Verderb zusätzliche Kosten verursachen. Sind die Bestände dagegen zu niedrig, entstehen Fehlmengen durch ausverkaufte Produkte, die zu Opportunitätskosten in Form von entgangenem Umsatz und Ertrag führen, bis hin zu einer Beeinträchtigung der Kundenzufriedenheit und Einkaufsstättenloyalität. Das Treffen korrekter Dispositionsentscheidungen auf der Basis einer Prognose des stochastischen Bedarfs im anonymen Markt ist somit von besonderer Bedeutung für den wirtschaftlichen Erfolg.

Hierbei kommen unterschiedliche Modelle und Verfahren der Bedarfsprognose zum Einsatz, die aufgrund ihrer unterschiedlichen Genauigkeit bei der Bedarfsvorhersage die Entscheidungsgüte entscheidend bestimmen. Die zugrundeliegende Arbeit von Crone (2009) untersucht die Verfahrensklasse der künstlichen Neuronalen Netze (NN) zur Prognose und erweitert diese mithilfe von asymmetrischen, betriebswirtschaftlichen Zielfunktionen der Entscheidungskosten zur simultanen Bedarfsprognose und Bestellmengenrechnung im Sinne einer integrierten Warendisposition. Die Prognosegenauigkeit und Dispositionskosten der NN werden anschließend in einer objektiven empirischen Vergleichsstudie mit etablierten Verfahren der Bedarfsprognose verglichen. Schwerpunkt dieses Beitrags ist die Darstellung der Problemstellung und Bedeutung der Bedarfsprognose im Handel um diese und zukünftige Forschungsanstrengungen zu motivieren, eine Einführung in die Verfahrensklasse der NN, und eine Darstellung ihrer möglichen Erweiterung zur simultanen Prognose und Bestellmengenrechnung.

1. Empirische Bedeutung der Bedarfsprognose

Die Prognose – im Sinne der stochastischen Vorhersage zukünftiger Ereignisse – liefert die Grundlage für alle betriebswirtschaftlichen Entscheidungen bei Unsicherheit, und repräsentiert einen zentralen Forschungsbereich des Operational Research (Fildes et al. 2008).

Handelsbetriebe wurden in den vergangenen Jahren mit zunehmenden Anforderungen an die Bedarfsprognose und Warendisposition konfrontiert (siehe u.a. Studien von (ECR Europe and Roland Berger 2003; Straube, Pfohl et al. 2005; KPMG and EHI 2006)). Das wachsende Preis-Leistungsbewusstsein und die sinkende Produkt-, Marken- und Einkaufsstättenloyalität der Konsumenten führte zu einem Zuwachs an Werbeaktivitäten, die eine Vervielfachung der Nachfrage um bis zu 600 % im Aktionszeitraum bewirken und bei einzelnen Nahrungsmitteln bereits bis zu 75 % des Gesamtumsatzes generieren. Neue Produkte werden in einer nie zuvor erreichten Geschwindigkeit und Menge eingeführt, so dass Sortimente im Handel trotz beschränktem Verkaufsraum kontinuierlich erweitert und durch Ein- und Auslistungen in immer kürzeren Zyklen erneuert werden. Als Resultat steigt die Volatilität und Komplexität der Nachfrage. Die Aufrechterhaltung der Lieferfähigkeit und der Termintreue bei wachsenden Nachfrageschwankungen erfordert eine Erhöhung der Warenbestände im Handel. Im Jahr 2005 investierten die einhundert größten Konsumgüterhersteller über 100 Mrd. US-Dollar in die Bestände in den handelsseitigen Fertigwarenlagern. Gleichzeitig stieg der Anteil unverkäuflicher Ware durch Beschädigung und Verderb auf 1,2 % des Umsatzes, was vermeidbaren Kosten von über 5 Mrd. US-Dollar pro Jahr entspricht. Dennoch sank die Warenverfügbarkeit in den Absatzstellen des Handels: 8 % aller neu eingeführten Produkte und 17 % der Produkte mit Sonderaktionen sind täglich ausverkauft. Dem europäischen Handel entgehen so 3 bis 4 % (etwa 4 Mrd. Euro) ihres Umsatzes pro Jahr, da Konsumenten fehlende Produkte nur zu 53 % durch alternative Packungsgrößen oder andere Marken substituieren, was die Bedarfsvolatilität weiter verstärkt.

Aus diesen Gründen gewinnen Ansätze zur Verbesserung der Bedarfsprognose immer mehr an Bedeutung. ECR Europe identifiziert die ungenaue Prognose von Standard- und Aktionswaren als einen Hauptverursacher von Ineffizienzen. Dabei können 35 bis 47 % aller Fehlmengen allein auf fehlerhafte Bedarfsprognosen in den Filialbestellungen zurückgeführt werden. Die Deutsche Bundesvereinigung Logistik nennt die »unzureichende Prognosefähigkeit« als Hauptursache für eine mangelnde Warenverfügbarkeit. Nach Einschätzung des US-Bundesverbands GMA beruhen von den 100 Mrd. US-Dollar an Beständen in Fertigwaren 50 % allein auf Fehlern in der

Absatzprognose. Dies ist darauf zurückzuführen, dass Prognosen der zukünftigen Bedarfsmengen und Bedarfszeitpunkte die Güte der Bestellmengenrechnung und somit die Güte der Warendisposition in zweierlei Hinsicht beeinflussen: (1.) Die Genauigkeit der Punktprognose bestimmt das Niveau der Bestellmenge und somit die Höhe der Abweichung vom tatsächlichen Bedarf. (2.) Die Prognosegenauigkeit vorangegangener Perioden bestimmt durch die Streuung historischer Prognosefehler zudem die Höhe der Sicherheitsbestände, welche Schwankungen in der Nachfrage und Ungenauigkeiten der Prognose ausgleichen sollen.

Aufgrund der Bedeutung der Prognosegenauigkeit für die Entscheidungsgüte der Warendisposition, auch in anderen Disziplinen, wurden in Wissenschaft und Praxis eine Vielzahl quantitativer Prognoseverfahren entwickelt. Trotz der besonderen Komplexität der Nachfrage im Handel, mit Ausreißern, Strukturbrüchen in Niveau und Trend, mehrfacher Saisonalität sowie externen, vielfach nichtlinearen Einflussfaktoren aus Kalendereffekten, Marketingpolitik und Wetter werden überwiegend einfache statistische Verfahren (u.a. der Exponentiellen Glättung) eingesetzt, so dass hier noch erhebliches Verbesserungs- und Forschungspotenzial besteht.

2. Neuronale Netze zur Bedarfsprognose

Als Alternative zu etablierten statistischen Verfahren können auch neuartige Ansätze der Computational Intelligence zur Bedarfsprognose verwendet werden. NN bilden eine Klasse semiparametrischer, datengetriebener und nichtlinearer mathematisch-statistischer Verfahren, deren Funktionsweise dem Vorbild der neurobiologischen Informationsverarbeitung in Nervenzellen und -verbänden des Menschen nachempfunden wurde. NN besitzen die Fähigkeit, jede beliebige lineare oder nichtlineare Funktion bis zu einem beliebigen Grad der Genauigkeit zu approximieren und auf neue Daten zu generalisieren, und so die erlernten Erklärungsmodelle im Sinne der Prognose zu nutzen (siehe u.a. Zhang et al. (1998)).

Die Prognose mit vorwärtsgerichteten, nicht-rekurrenten NN umfasst die Vorhersage einer abhängigen Variablen \hat{y} aus zeitlich verzögerten Realisationen derselben y_{t-n} , i unabhängigen Erklärungsvariablen x_i mit metrischem, ordinalen oder nominalen Skalenniveau sowie zeitlichen Verzögerungen davon, $x_{i,t-n}$. Somit bieten NN eine Vielzahl von Freiheitsgraden in der Modellierung zur Prognose, einschließlich der Bildung linearer und nichtlinearer Modelle der Zeitreihenanalyse der abhängigen Variablen, $\hat{y}_{t+1} = f(y_t, y_{t-1}, \dots, y_{t-n+1})$, Regressionsmodellen der Kausalanalyse im Sinne von $\hat{y} = f(x_1, x_2, \dots, x_i)$, als auch Interventionsmodelle und Transferfunktionsmodelle, $\hat{y}_{t+1} = f(y_t, y_{t-1}, \dots, y_{t-n+1}, x_i)$ usw. Sie bilden somit nahezu das gesamte Spektrum dynamischer Regressionsmodelle ab. Die Prognose mit NN erfolgt meist als nicht-lineares autoregressives AR(p)-Modell der Zeitreihenanalyse unter Verwendung eines Multi-layer Perceptrons (MLP), der populärsten Architektur von NN. Zu jedem Zeitpunkt t wird eine Prognose \hat{y}_{t+h} für den Zeitpunkt $t+h$ (mit dem Prognosehorizont h vielfach als $h=1$), unter Verwendung der p Beobachtungswerte $y_t, y_{t-1}, \dots, y_{t-p+1}$ der vorherigen Zeitperioden $t, \dots, t-p+1$ erstellt. Die Anzahl der autoregressiven Terme p entspricht dann der Anzahl der Eingabeneuronen, und die Anzahl der abhängigen Variablen bzw. Prognosehorizonte $t+1, \dots, t+h$ den Ausgabeneuronen des NN. Aus Gründen der Übersichtlichkeit wird die Architektur anstelle der mathematischen Formulierung als gerichteter Graph des MLP dargestellt, siehe Abb.1.

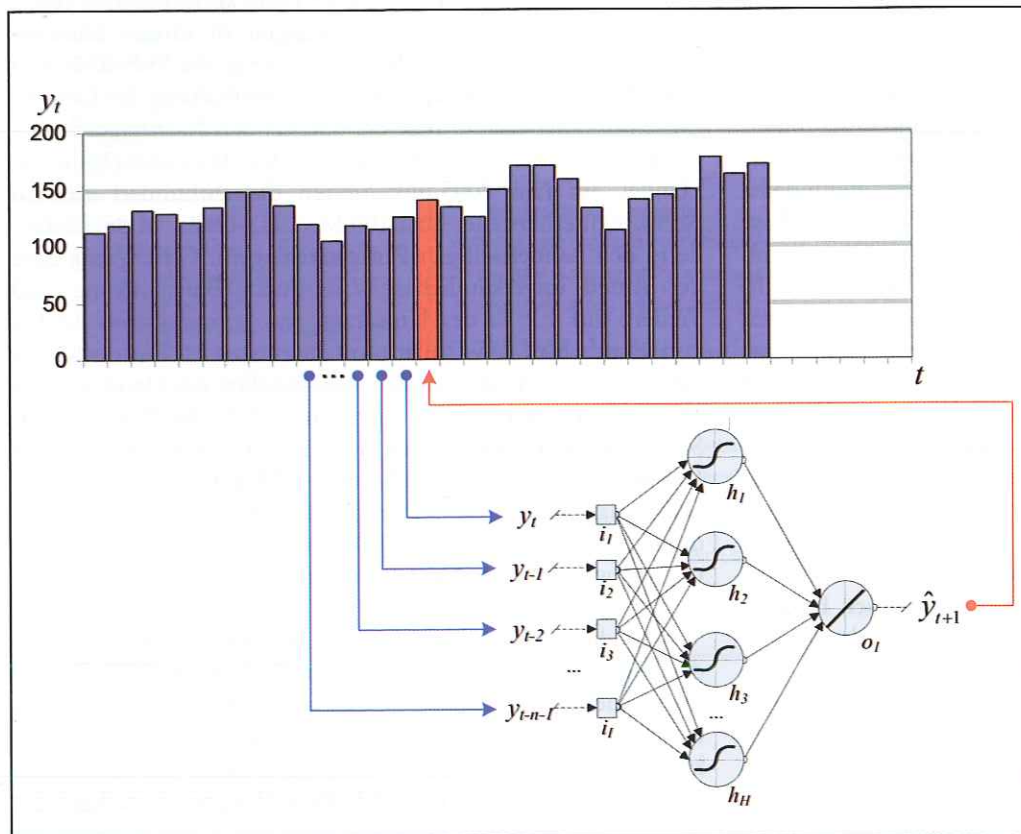


Abbildung 1: Autoregressives MLP zur Zeitreihenprognose, mit p input Neuronen, m versteckten Neuronen und h Ausgabeneuronen (jeweils mit trainierbaren Schwellwerten) und zwei Schichten trainierbarer Verbindungsgewichte

Innerhalb der Netzwerkarchitektur kann die Funktionalität eines einzelnen Neurons u_i im Kontext statistischer Prognoseverfahren interpretiert werden. Jedes Neuron berechnet die Summe der Eingabewerte x_1, x_2, \dots, x_n und einem Schwellwert θ_j gewichtet mit n Verbindungsgewichten $w_{0j}, w_{1j}, \dots, w_{nj}$ (die im Training als freie Parameter bestimmt werden), welche durch eine Aktivierungsfunktion $f_{act}(\cdot)$ transformiert und ausgegeben wird. Für eine lineare f_{act} berechnet ein einzelnes Neuron somit eine konventionelle lineare Regression, mit $w_{ij} = \beta_{ij}$ und $\theta_j = \beta_0$, bzw. im Fall von n zeitlich verzögerten Variablen eine lineare Autoregression $AR(p)$, für eine nichtlineare f_{act} entsprechend eine nichtlineare (logistische) Regression oder Autoregression $NAR(p)$ sowie unter Berücksichtigung kausaler Variablen ein Interventionsmodell $NARX(p)$ usw. Betrachtet man die Familie autoregressiver Prozesse wird zudem deutlich, dass bei geeigneter Auswahl der zeitlich verzögerten Variablen und der Verbindungsgewichte auch Naive Verfahren, gleitende Mittelwerte, Exponentielle Glättung einschließlich trendbehafteter und saisonaler Varianten approximativ durch ein einzelnes Neuron abgebildet werden können. Die Kombination von Neuronen in einem Netzwerk bildet somit ein gewichtetes Ensemble von (nicht-)linearen Regressionsmodellen.

Durch die Berücksichtigung von Rückkoppelung der Ausgabe zu Neuronen in vorgelagerten Schichten lassen sich ferner rekurrente NN bilden, die auch zeitlich verzögerte Prognose-

fehler im Sinne von linearen und nichtlinearen moving average $MA(q)$ -Prozessen effizient abbilden können (Zimmermann and Neuneier 2000).

Trotz ihrer theoretischen Fähigkeiten sind NN bislang nicht als Prognoseverfahren in der Praxis etabliert, im Gegensatz zu ihrer breiten Akzeptanz als Verfahren der Klassifikation. Die Einsatzfähigkeit von NN zur Prognose wurde seit Beginn der 1990er-Jahre in über 5.000 Publikationen in diversen Anwendungsbereichen experimentell demonstriert. Ein Großteil dieser Studien lässt jedoch aufgrund von mangelhaften Experimentaufbauten (einzelne oder wenige Zeitreihen, einzelne Zeitreihenursprünge, ungeeignete Fehlermaße, keine Analyse der Bedingungen unter denen NN besser/schlechter prognostizieren, siehe Küsters (2006)), unzureichender Dokumentation die eine Replikation verhindert sowie dem mangelndem Vergleich der Ergebnisse mit der Genauigkeit etablierter statistischer Verfahren nur eingeschränkt valide und reliable Erkenntnisse zu (Adya and Collopy 1998). Dies verdeutlicht jedoch gleichzeitig das Potenzial zukünftiger Untersuchungen mit wohlüberlegten Experimentdesigns.

Die vorliegende Arbeit untersucht daher u.a. auch die Anwendung von NN als Prognoseverfahren in einem experimentellen Vergleich auf archetypischen, synthetischen Zeitreihen nach der Pegels-Gardner-Klassifikation sowie einer repräsentativen Auswahl empirischer Zeitreihen.

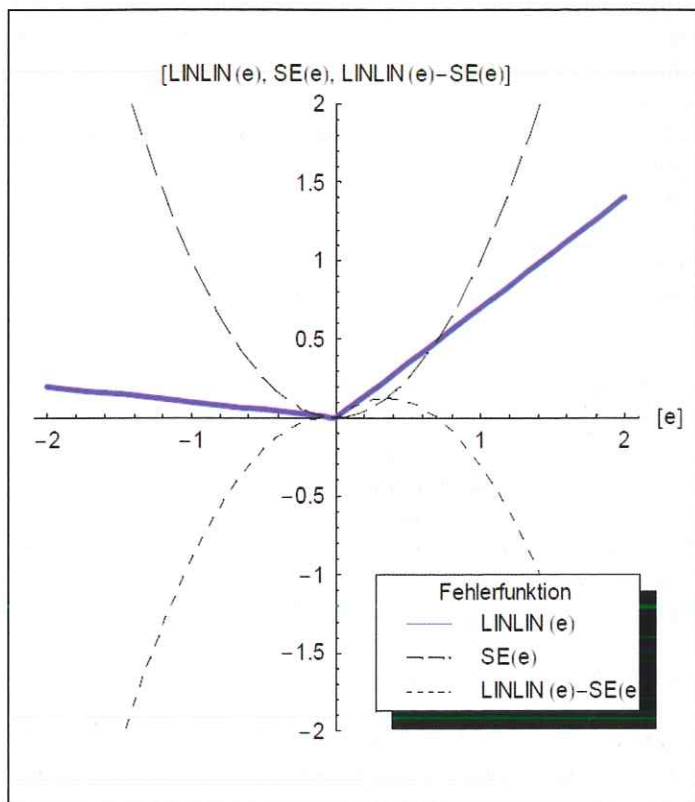


Abbildung 2: Asymmetrische LINLIN(e)-Kostenfunktion im Vergleich zum quadratischen Fehler $SE(e)$ und die Differenz beider Funktionen $LINLIN(e)-SE(e)$

3. Künstliche Neuronale Netze zur Warendisposition

Wie viele lineare statistische Verfahren der Regression finden NN adäquate Parameter (die Verbindungsgewichte w_{ij}) durch die Minimierung einer quadratischen Zielfunktion (SE , im Sinne des l_2 -loss etablierter Kleinste-Quadrate-Schätzer), welche durch die Bewertung der Fehlerkosten die Zielsetzung und somit das Verhalten des NN als Prognoseverfahren wesentlich bestimmt.

Aufgrund der direkten Einbindung der Zielfunktion in den populären Backpropagation-Algorithmus wurde die Nutzung abweichender statistischer Fehlermaße zur Prognose mit NN nur vereinzelt vorgeschlagen, so zum Beispiel die Funktion der absoluten Fehler (l_1 -loss) sowie der ϵ -insensitiven Verlustfunktion zur robusten Prognose, die beide symmetrisch um Null verlaufen. Da in der betriebswirtschaftlichen Praxis die Kosten fehlerhafter Entscheidungen vielfach quantifizierbar sind, können diese unmittelbar als Ziel- bzw. Kostenfunktion modelliert werden. In der Warendisposition im Handel steigen die Kosten einer Überprognose oder Unterprognose in der Regel nicht symmetrisch und nicht quadratisch mit dem Betrag des Prognosefehlers: Fehlmengen durch zu niedrige Prognosen führen zu höheren Kosten als Überbestände durch zu hohe Prognosen. Können diese Kosten geschätzt werden, erlaubt dies die Bildung von asymmetrischen, nichtquadratischen Kostenfunktionen (ACF), die als asymmetrische Zielfunktionen die

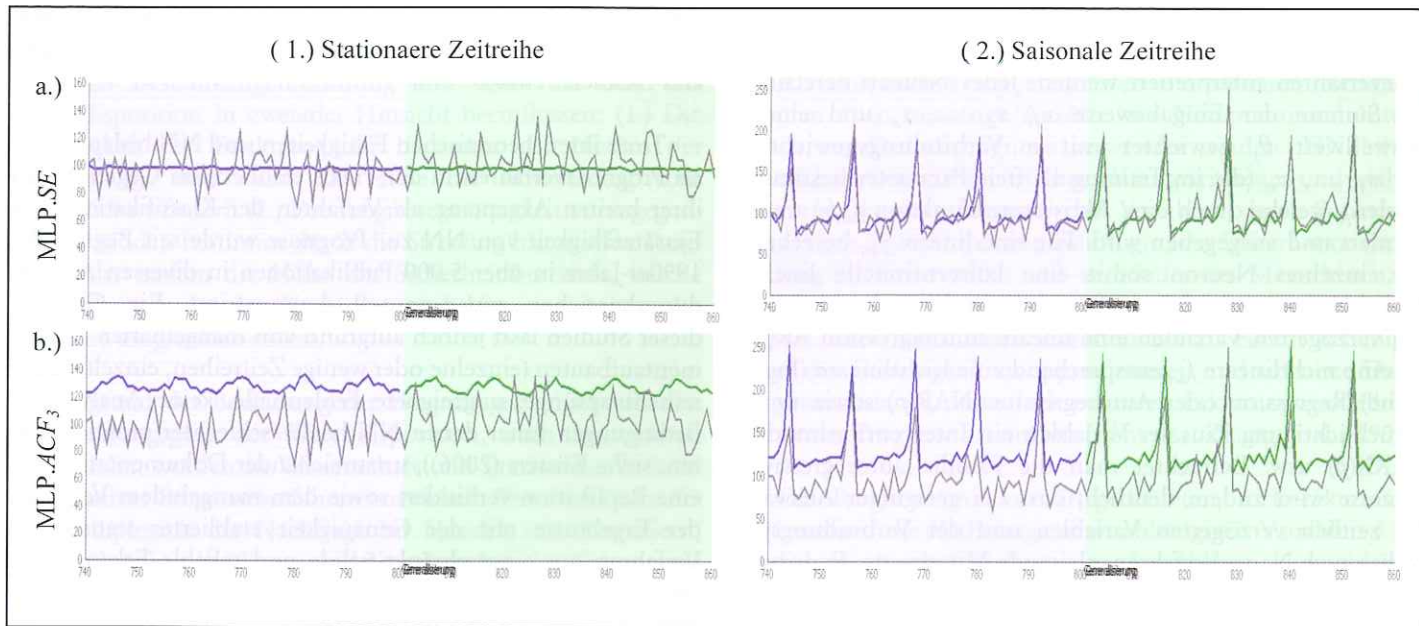


Abbildung 3: Stationäre (1) und saisonale (2) Originalzeitreihe mit den Prognosen des MLP.SE (a) und des MLP.ACF₃ (b)

tatsächlichen betrieblichen Entscheidungskosten von positiven und negativen Prognosefehlern aus der Bestandsentscheidung der Warendisposition repräsentieren (Granger 1969), siehe Abb. 2.

Der Grad der Asymmetrie wird durch das Verhältnis von Entscheidungskosten aus zu niedrigen Prognosen, c_u , und zu hohen Prognosen, c_o , bestimmt, welches auch den kostenoptimalen α -Servicegrad für das Newsboy-Modell der Bestellmengenrechnung bei einperiodigem, stochastischem Bedarf angibt, $c_o / (c_o + c_u)$.

Die Verwendung von ACF im Training von NN erfordert lediglich eine Erweiterung des Backpropagation-Algorithmus, und erlaubt eine direkte Überprüfung der Auswirkungen in einem empirischen Vergleich. Hierfür wird ein MLP mit konstanter Architektur (13 Eingangs-, zwei versteckte und ein Ausgabe-Neuron) mit unterschiedlichen Zielfunktionen auf synthetischen Zeitreihen trainiert. Abb. 3 zeigt die Ausgabe des MLP parametrisiert auf einer stationären Zeitreihe (1) und einer saisonalen Zeitreihe (2) unter Verwendung einer quadratischen Fehlerfunktion (entspricht einem 50 % α -Servicegrad) (a) und einer ACF (entspricht einem 99 % α -Servicegrad) (b), und erlaubt so eine grafische Interpretation ihrer Auswirkung auf die Prognose.

Ein MLP, das mit dem quadratischen Fehler SE trainiert wird, erkennt Niveau und ggf. saisonales Muster und prognostiziert den Erwartungswert valide und reliabel auf den Trainings- und Testdaten, sowohl für die stationäre als auch saisonale Zeitreihe (bei leichtem Überlernen). Im Gegensatz dazu erhöht das MLP, das mit der ACF trainiert wird, seine Netzwerkausgabe, und prognostiziert so entsprechend der Asymmetrie der Kosten

höhere Werte für die stationäre und saisonale Zeitreihe. Diese Ausgabe entspricht nicht länger dem Mittelwert der Wahrscheinlichkeitsfunktion, sondern einem Quantil auf der kumulativen Verteilungsfunktion der Prognosefehler, das durch die Dispositionskosten der Zielfunktion spezifiziert wurde. Somit verändert das MLP seine Ausgabe von einer Prognose des fehlerminimalen Erwartungswertes hin zu einer direkten Vorhersage des kostenminimalen Warenbestandes für die zugehörige Bedarfsstruktur. Dies entspricht dem Erwartungswert mit den minimalen Entscheidungskosten für einen zukünftigen Dispositionszeitpunkt. Unter Verwendung einer asymmetrischen Kostenfunktion aus den tatsächlichen Entscheidungskosten kann die Funktionsweise des NN somit zu einem integrierten Entscheidungsmodell der simultanen Bedarfsprognose und Bestandsberechnung in der Warendisposition erweitert werden.

Abb. 4 belegt das abweichende Verhalten eines MLP für vier Kostenfunktionen mit zunehmender Asymmetrie als Boxplots der Servicegrade über eine Vielzahl von Zeitreihen. In Abhängigkeit von der Kostenfunktion erzielt das MLP die korrespondierenden Servicegrade von $SE=50\%$, und $ACF=\{85\%, 95\%, 99\%\}$, und prognostiziert somit jene Quantile auf der Verteilungsfunktion, die mit der jeweiligen Asymmetrie der Kosten c_u und c_o korrespondieren. Das Verhalten zeigt sich konsistent über Trainings- und Testdaten, und bestätigt so die Fähigkeit von MLPs zur Approximation und Generalisierung mit ACF ohne A-priori-Annahmen oder statistische Tests über die Verteilungsfunktion der Bedarfe bzw. der Prognoseressiduen oder eine iterative Modellbildung durch Experten oder Expertensysteme.

Die Ergebnisse sind darüber hinaus konsistent über eine repräsentative Auswahl synthetischer Zeitreihen mit Verteilungsfunktionen von normal-, poisson- und negativ-binomialverteilt.

tem Bedarf sowie unterschiedlicher Streuung der Residuen. Im Vergleich zu statistischen Prognoseverfahren und etablierten Expertensystemen zeigen MLPs mit *ACF* statistisch signifikant niedrigere Dispositionskosten und somit effizientere Warenbestände als bei konventioneller Berechnung von Sicherheitsbeständen. Ferner bestätigen sich diese Ergebnisse in einem umfassenden Verfahrensvergleich von zeitreihen- und kausalanalytischen Verfahren auf empirischen Zeitreihen des Absatzes an Warenautomaten, als besondere Form des stationären Einzelhandels (für eine detaillierte Analyse siehe Crone (2009)).

4. Zusammenfassung und Ausblick

Die zugrundeliegende Dissertation untersucht die Verfahrensklasse der NN zur Prognose und erweitert sie durch die Verwendung von *ACF* erstmalig zur simultanen Bedarfsprognose und Bestellmengenrechnung im Sinne einer integrierten Entscheidungsfindung der Warendisposition, was in der wissenschaftlichen Literatur bisher vernachlässigt wurde. Das NN prognostiziert unter Verwendung der *ACF* einen kostenminimalen Warenbestand als beliebiges Quantil einer Verteilungsfunktion anstelle eines fehlerminimalen Bedarfswertes, wobei ex ante keine Verteilungsannahmen wie für statistische Verfahren erforderlich sind. Die Fähigkeit der NN wird in einer Vergleichsstudie auf synthetischen Zeitreihen ermittelt und anschließend in einem objektiven empirischen Verfahrensvergleich mit etablierten Verfahren der Bedarfsprognose bestätigt.

Die dargestellte Funktionalität und die Güte der MLPs mit *ACF* rechtfertigen ihren Einsatz zur Disposition auf Zeitreihen bei stationärem sowie instationärem Bedarf unterschiedlicher Streuung und beliebiger Verteilungsformen. In der empirischen Anwendung kann somit eine einzige Konfiguration eines MLP sortimentsübergreifend zur effizienten Disposition eingesetzt

werden. Ferner werden die Teilschritte der Bedarfsprognose und Bestellmengenrechnung zur Bestimmung kostenminimaler Warenbestände integriert, und die Modellbildung somit näher an die betriebliche Realität geführt. Die Übertragung der Erkenntnisse auf andere Betriebsformen und Branchen in Industrie und Handel liegt aufgrund der übergreifenden Problemstellung der Disposition nahe. Die Erweiterung der Theorie der NN auf das Erlernen alternativer Zielfunktionen ermöglicht ferner neuartige Anwendungen über die Disziplinen der Betriebswirtschaftslehre und Prognostik hinaus. In der Zukunft sind jedoch zunächst die Erkenntnisse dieser Arbeit durch Vergleichsstudien auf repräsentativen Stichproben von empirischen Zeitreihen kritisch zu überprüfen. Erst dies erlaubt den Rückschluss auf eine allgemeine Einsatzfähigkeit von NN mit *ACF* in der betrieblichen Entscheidungsunterstützung.

Danksagung

Ich bedanke mich herzlich bei der GOR und der Jury für die Auszeichnung, und für die Gelegenheit meine Arbeit bei der EURO'09 präsentieren zu dürfen.

Literatur

- [1] Adya, M. and F. Collopy (1998). *How effective are neural networks at forecasting and prediction? A review and evaluation.* Journal of Forecasting 17(5-6): 481-495.
- [2] Crone, S. F. (2009). *Neuronale Netze zur Prognose und Disposition im Handel.* Wiesbaden, Gabler.
- [3] ECR Europe and Roland Berger (2003). *ECR - Optimal Shelf Availability.* ECR Europe: 65.
- [4] Granger, C. W. J. (1969). *Prediction With A Generalized Cost Of Error Function.* Operational Research Quarterly 20(2): 199-208.
- [5] KPMG and EHI (2006). *Trends im Handel 2010.* Consumer Markets & Retail Reports. KPMG. Köln: 78.
- [6] Küsters, U. (2006). *Evaluation, Kombination und Auswahl betriebswirtschaftlicher Prognoseverfahren.* Prognoserechnung. P. Mertens and S. Rässler. Heidelberg, Physica-Verlag: 367-404.
- [7] Straube, F., H.-C. Pfohl, et al. (2005). *Trends und Strategien in der Logistik: ein Blick auf die Agenda des Logistik-Managements 2010.* Hamburg, Dt. Verkehrs-Verl.
- [8] Zhang, G., B. E. Patuwo, et al. (1998). *Forecasting with artificial neural networks: The state of the art.* International Journal of Forecasting 14(1): 35-62.
- [9] Zimmermann, H.-G. and R. Neuneier (2000). *Modeling dynamical systems by recurrent neural networks.* Data Mining II. Southampton, Wit Press. 2: 557-566.

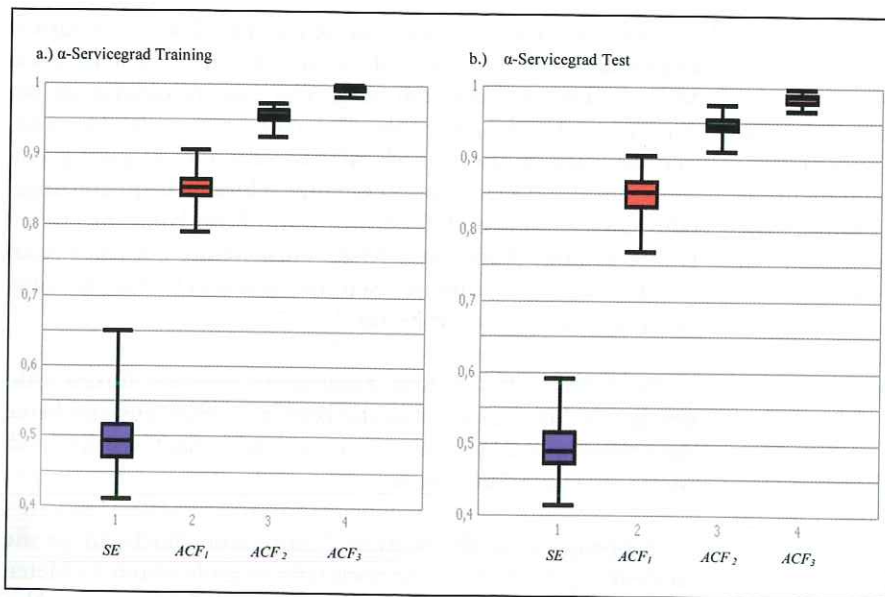


Abbildung 4: Verteilung der α -Servicegrade über alle MLP auf Trainings- (a) und Testmenge (b) für die Zielfunktionen SE , ACF_1 , ACF_2 und ACF_3

