OR Spektrum (1993) 15:21–30

OR Spektrum

Eine Produktions-/Lagerhaltungspolitik bei stochastischer Nachfrage im Mehrproduktfall

Hermann Jahnke

Institut für Logistik und Transport, Universität Hamburg, Von-Melle-Park 5, W-2000 Hamburg 13

Eingegangen am 12. März 1992 / Angenommen am 21. Oktober 1992

Zusammenfassung. Eine Lagerzielmengenpolitik für den Fall intermittierender Fertigung mehrerer Sorten auf einer Anlage bei stochastischer Nachfrage wird durch ein Warteschlangenmodell abgebildet. Für die sich ergebenden Kostenfunktion wird eine Approximation vorgeschlagen. Die Bestimmung optimaler Losgrößen und Lagerzielmengen wird untersucht.

Summary. A queueing model is presented for a base-stock policy in the multi-product/one-machine case with stochastic demand. The resulting cost function is approximated using a Browian Motion. The numerical optimization of lot-sizes and base-stock levels is discussed.

Schlüsselwörter: Produktionsplanung, Losgrößen, Lagerhaltung, Warteschlangen

Keywords: Production planing, Rot-sizing, inventory, queueing

1. Einleitung

In der vorliegenden Arbeit soll eine Lagerzielmengen- oder Base-Stock-Politik bei losweiser, geschlossener Fertigung mehrerer Produkte (Sorten) mit fester Losgröße auf einer Anlage in kontinuierlicher Zeit untersucht werden. Es wird eine beschränkte Produktionsgeschwindigkeit betrachtet, d. h. die Bearbeitung der Lose nimmt jeweils eine endliche positive Zeit in Anspruch. Die Lose werden nach dem jeweiligen Herstellungsende in ein Lager eingestellt, aus dem die in zufälligen Zeitpunkten auftretende Nachfrage befriedigt wird. Mögliche Fehlmengen werden vorgemerkt.

Die Literatur, die diese komplexe Produktionssituation zum Gegenstand ihrer Untersuchung macht, ist nicht sehr umfangreich. Zu erwähnen sind vor allem die Arbeiten von Williams [13] und Zipkin [14]. Sie kombinieren Modelle der Lagerhaltungs- und der Warteschlangentheorie: Fertigungsaufträge vom Umfang der Losgröße werden ausgelöst, sobald der Lagerbestand des betreffenden Produktes eine Meldemenge erreicht oder unterschreitet. Die Fertigungsaufträge ordnen sich in eine Warteschlange vor der Fertigungsanlage ein, werden der Reihe nach bearbeitet und anschließend in ihr Lager transportiert, wobei die Bearbeitungszeiten – wie Williams und Zipkin annehmen – stochastisch sind. Die zu optimierende Kostenfunktion ist an die Zielgrößen der Lagerhaltungstheorie angelehnt, wobei die Rolle der (zufälligen) Lieferzeit von der aus dem Warteschlangenmodell ermittelten Systemzeit (Warte- plus Bearbeitungszeit) der Aufträge bzw. Lose übernommen wird. Beide Autoren approximieren den Ankunftsstrom der Aufträge durch einen Poissonprozeß, was gerechtfertigt erscheint, wenn eine große Anzahl von Produkten betrachtet wird, die eine niedrige Auflagehäufigkeit und damit relativ große Lose haben.

Der skizzierte Ansatz für die Produktionssituation macht einen wesentlichen Unterschied zu den Modellen der Lagerhaltungstheorie deutlich: Die Durchlaufzeit eines Auftrages, also die Zeit, die zwischen der Erteilung eines Fertigungsauftrages und dem Ende seiner Bearbeitung verstreicht, ist über die Rüst- und die Bearbeitungsdauer von der Losgröße der betrachteten Sorte abhängig. Bei endlicher Produktionsgeschwindigkeit und knappen Kapazitäten wird darüberhinaus das Warten auf das Freiwerden der Anlage einen entscheidenden Bestandteil dieser Zeitdauer ausmachen. Die Wartezeit hängt aber von der Anzahl zuvor erteilter Aufträge der anderen Produkte, also auch von deren Losgröße und ihrer stochastischen Nachfrage ab. Dieser Warteschlangeneffekt wird von den Modellen der Lagerhaltungstheorie im allgemeinen nicht betrachtet. Für die einzelnen Produkte werden aber sowohl die Lagerhaltungs- und Fehlmengenkosten als auch der Servicegrad des Lagers von der Durchlaufzeit der Fertigungsauträge beeinflußt.

Die geeigneten Mittel, den Warteschlangeneffekt in ein Modell für die beschriebene Produktionssituation einzubeziehen, liefert die Warteschlangetheorie. Sie bildet hier daher, wie in den Arbeiten von Williams und Zipkin, die Basis der Abbildung der Produktionsvorgänge.

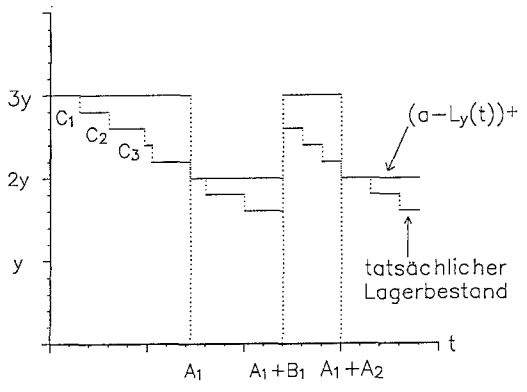

Abb. 1. Entwicklung des tatsächlichen Lagerbestandes und der Modellgröße $(a-L_y(t))^+$, die unten definiert wird

Im Unterschied zu den von Williams und Zipkin vorgeschlagenen Ansätzen werden bei der in dieser Arbeit untersuchten Lagerzielmengenpolitik Fertigungsaufträge eines festen, sortensspezifischen Umfanges ausgelöst, sobald der Absatz eines entsprechenden Loses aus dem Lager vollständig abgeschlossen, d. h. eine über einen gewissen Zeitraum kumulierte Nachfrage in Höhe eines Loses aufgetreten ist. Die Bearbeitung von Fertigungsaufträgen eines Produktes wird eingestellt, sobald der aktuelle Lagerbestand eine gewisse Höchstmenge – die Lagerzielmenge – erreicht.

Motiviert ist die Untersuchung dieser Produktions-/Lagerhaltungspolitik durch die Beobachtung, daß die Lagerzielmengenpolitik bei diskreter Periodeneinteilung, unbeschränkter Produktionsgeschwindigkeit und stochastischer Nachfrage im Einproduktfall optimal ist (vgl. [6], S. 64–68). Im Gegensatz zu diesen Rahmenbedingungen wird in der vorliegenden Arbeit angenommen, daß die Produktionsgeschwindigkeit endlich, d. h. die Kapazität beschränkt ist. Zum anderen wird auf die diskrete Zeiteinteilung verzichtet. Die Einteilung der Zeit in Perioden resultiert nämlich in der Regel weniger aus den physischen Herstellungsvorgängen oder dem Strom der Nachfrage selbst. Beide kann man im allgemeinen als Prozesse mit kontinuierlicher Zeit auffassen. Die in den Modellen und der Praxis der Produktionsplanung häufige Periodisierung wird eher durch die verwendeten Planungsverfahren und die Notwendigkeit verursacht, bei Datenunsicherheit Schätzungen in bestimmten Zeitabständen zu erneuern und anschließend in die Planung einzuarbeiten. Die hier eingesetzten warteschlangentheoretischen Instrumente benötigen aber keine Periodisierung der Zeit. Der revidierende Einsatz des vorgeschlagenen Ansatzes kann zwar in der praktischen Anwendung sinnvoll sein, da in der Regel die benötigten Kenntnisse über die Parameter der beteiligten Verteilungen (Erwartungswerte, Varianzen) fehlen und durch Schätzungen ersetzt werden müssen; für die Entwicklung des Ansatzes wird hier jedoch die Kenntnis der Verteilungsparameter unterstellt.

Mit der Übertragung von Produktionspolitiken auf eine Situation mit endlicher Produktionsgeschwindigkeit und kontinuierlicher Zeit, die dann im wesentlichen mit warteschlangentheoretischen Hilfsmitteln untersucht wird, beschäftigen sich auch Gaver ([2], Lagerzielmengenpolitik), Altiok [1] und die dort angegebenen Literaturstellen ((*s*, *S*)-Politiken). In diesen Arbeiten werden meistens der Einsortenfall mit einer Fertigungsauftragsgröße von einer Mengeneinheit und ein poissonscher Nachfrageprozeß unterstellt. Diese Einschränkungen sollen hier aufgehoben werden.

In Abschn. 2 wird zunächst ein warteschlangentheoretisches Modell für den Einsortenfall dargestellt. In Abschn. 3 wird analog zu diesem Ansatz die Kostenfunktion im Mehrsortenfall hergeleitet. Sie enthält Kenngrößen von im allgemeinen unbekannten Verteilungen, die auf geeignete Weise zu approximieren sind (Abschn. 4). In Abschn. 5 finden sich Überlegungen zur optimalen Wahl der Lagerzielmengen und der Losgrößen der Produkte.

2. Die Kostenfunktion im Einsortenfall

Für die Modellierung des stochastischen Nachfragestromes wird angenommen, daß die zwischen je zwei Nachfragevorgängen vergehenden Zeiten eine Folge von zufälligen, voneinander unabhängigen und der gleichen Verteilung gehorchenden Zufallsvariablen $C_1, C_2, \ldots$ mit Erwartungswert $0 < E(C_i) = E(C_1) = 1/\lambda < \infty$ (für alle $i = 1, 2, \ldots$) bilden. In den Nachfragezeitpunkten wird jeweils eine Mengeneinheit des Produktes nachgefragt. Nachfrage, die auf ein leeres Lager trifft, wird vorgemerkt und später befriedigt (Vormerkfall), also zu eventuell höheren Kosten „nachgeliefert".

Die Lagerzielmengenpolitik arbeitet mit einer festen Losgröße von y Mengeneinheiten. Bevor ein Fertigungsauftrag erteilt wird, werden jeweils y Nachfrageeinheiten gesammelt, d. h. entweder aus dem Lager befriedigt oder vorgemerkt. Zwischen den Zeitpunkten, zu denen der Absatz des $(i-1)$-ten bzw. des i-ten Loses abgeschlossen und deshalb jeweils ein Produktionsauftrag über y Mengeneinheiten erteilt wird, liegt die Summe von y Zwischennachfragezeiten, also $A_i = \sum_{j=(i-1)y+1}^{iy} C_j$ Zeiteinheiten. Die Erwartungswerte dieser Größen sind $E(A_i) = E(A_1) = y/\lambda$ (für alle $i = 1, 2, \ldots$).

Die Bearbeitungszeit B_1 des ersten Loses setzt sich aus der Rüstzeit von r Zeiteinheiten und der reinen Bearbeitungszeit von y/v Zeiteinheiten (die Produktionsgeschwindigkeit betrage v Mengeneinheiten pro Zeiteinheit) zusammen: $B_1 = r + y/v$. Die Bearbeitungszeit der anderen Lose ergibt sich analog. Nach dem Ende der Bearbeitung wird das Los komplett in das Lager eingestellt. Es wird eine geschlossene Fertigung unterstellt, bei der das Los erst ab dem Zeitpunkt seines Bearbeitungsendes zur Nachfragebefriedigung zur Verfügung steht. Die geschilderte Lagerbestandsentwicklung entspricht dem „tatsächlichen Lagerbestand" des in Abb. 1 dargestellen Lagerbestandspfades für den Fall einer Losgröße von fünf Mengeneinheiten und eines Anfangbestandes von drei Losen.

Die durch die Komplettierung des Absatzes eines Loses ausgelösten Produktionsaufträge bilden Ankünfte in einem $G/G/1$-Warteschlangensystem mit Zwischenankunftszeiten $A_1, A_2, \ldots$, (deterministischen) Bediendauern $B_1, B_2, \ldots$ und der Verkehrsintensität $\varrho(y) =$

$(\lambda r/y)+(\lambda/v)$ ([5], S. 183). Die Verkehrsintensität läßt sich bei gegebener Losgröße als der Zeitanteil interpretieren, der zur Befriedigung der auftretenden Nachfrage für Rüst- oder Herstellungsvorgänge verwendet werden muß. Denn λ/v ist der Anteil an einer Zeiteinheit für die reine Bearbeitungszeit, die benötigt wird, um eine Produktion in Höhe der Nachfragerate aufrecht zu erhalten. Dagegen ist λ/y die Rate, mit der durch Lagerabgänge Fertigungsaufträge ausgelöst werden, so daß der erste Term den Rüstzeitanteil an einer Zeiteinheit wiedergibt.

$\varrho(y)$ ist zugleich ein Auslastungsgrad. Denn die Anlage steht still (Rüstzeiten zählen nicht als Stillstandszeiten), falls im Zeitpunkt t die Warteschlangenlänge $L_y(t)$ – d. h. die Anzahl erteilter und noch nicht vollständig bearbeiteter Aufträge – verschwindet, also $L_y(t)=0$. Nun ist aber $\varrho(y)=1-P(L_y=0)$, wobei $P(L_y=0)$ die Wahrscheinlichkeit für den Zustand 0 unter der Grenzverteilung des Warteschlangenlängenprozesses $\{L_y(t), t\geq 0\}$ ist. Die Warteschlangenlänge ist ein regenerativer stochastischer Prozeß ([5], S. 180 und 183), so daß $P(L_y=0)$ zugleich der Anteil der erwarteten Stillstandszeiten während des ersten Regenerationsintervalles an der erwarteten Länge dieses Intervalles ist ([5]), S. 181). $\varrho(y)$ ist folglich der Anteil der erwarteten Betriebszeit der Anlage, also der erwartete Auslastungsgrad im Regenerationsintervall. Wächst die Losgröße, so werden Produktionsaufträge seltener erteilt, und der Auslastungsgrad der Anlagen sinkt.

Sei S der zufällige Zeitpunkt, in dem das erste Regenerationsintervall $[0, S)$ endet und das zweite beginnt, d. h. der Zeitpunkt, in dem nach dem Ende der ersten Betriebsperiode ein Produktionsauftrag auf eine leerstehende Anlage trifft. Für S wird eine nichtarithmetische Verteilung unterstellt, so daß die erwähnte Grenzverteilung existiert, wenn nur Losgrößen mit Auslastungsgraden $\varrho(y)<1$ betrachtet werden ([5], S. 183). Es wird gefordert, daß zu Anfang des Betrachtungszeitraumes ein Produktionsauftrag erteilt wird, d. h. eine erste Ankunft stattfindet.

Im weiteren werden folgende Bezeichnungen benutzt: $(x)^+$ steht für das Maximum von x und 0; $\mathbf{1}(\cdot)$ für die Indikator- oder charakteritische Funktion – d. h. $\mathbf{1}(X\leq x)$ ist gleich 1, falls $X\leq x$ gilt und sonst gleich 0 –; $F_L(\cdot; y)$ für die Verteilungsfunktion der Grenzverteilung und $E(L_y)$ für ihren Erwartungswert.

Im Vormerkfall geht keine Nachfrage endgültig verloren, der erwartete (Brutto-) Deckungsbeitrag pro Zeiteinheit ist in Bezug auf die Losgröße und das angestrebte Lagerniveau eine Konstante. Es ist unter der Zielsetzung Gewinnmaximierung daher optimal, die Losgröße y und die Lagerzielmenge a (gemessen in der Anzahl Lose; in der in Abb. 1 dargestellten Situation ist a beispielsweise gleich Drei) so zu wählen, daß sie die Summe aus Fehlmengen-, Lagerhaltungs- und Rüstkosten minimieren. Folglich soll der Erwartungswert dieser Kosten im ersten Regenerationsintervall hergeleitet und auf eine Zeiteinheit bezogen werden ($K(a; y)$).

Um Lagerhaltungs- und Fehlmengenkosten zu ermitteln, werden zunächst ein Zeitpunkt t, in dem der Absatz eines Loses gerade abgeschlossen und ein Fertigungsauftrag erteilt wird, und eine angestrebte Lagerzielmenge von a Losen betrachtet. Geht man beispielsweise davon aus, daß im Anfangszeitpunkt a Lose auf Lager liegen, ist die in t im Lager befindliche Anzahl von Losen gleich der Lagerzielmenge abzüglich der erteilten Fertigungsaufträge $L_y(t)$, solange $L_y(t)$ nicht größer als a ist. Diese Losanzahl, die in die Berechnung der Lagerhaltungskosten einfließt, ist also $(a-L_y(t))^+$. Eine negative Differenz aus Lagerzielmenge und Auftragszahl in t entspricht der Fehlmenge im Zeitpunkt t der Auftragserteilung. Die Anzahl von Losen, für die bis t Nachfrage aufgetreten ist und die vorgemerkt werden mußte, ist $(L_y(t)-a)^+$. Die Fehlmenge in t ist gleich dem Produkt aus der Anzahl der Fehllose und der Losgröße, also gleich $y(L_y(t)-a)^+$. Die für die Auftragserteilungszeitpunkte hergeleiteten Formeln für Lagerbestand und Fehlmenge werden in diesem und dem nächsten Abschnitt auch für die anderen Zeitpunkte übernommen, um die zugehörigen Kosten ermitteln zu können. Diese Kostenermittlung ist daher relativ grob. Beispielsweise wird der Lagerbestand nur ungenau erfaßt, denn zwischen zwei aufeinanderfolgenden Auftragserteilungszeitpunkten wird ein Los gerade abgesetzt, ist also während dieser Zeitspanne nicht dauernd vollständig im Lager vorhanden (vgl. in Abb. 1 die Differenz zwischen tatsächlichem Lagerbestand und der entsprechenden Modellgröße). Eine Korrektur des Lagerbestandes erfolgt im Modell aber erst, wenn der Absatzvorgang beendet und ein Fertigungsauftrag erteilt wird. Insofern überschätzt das Modell den tatsächlichen Lagerbestand. Auf eine verfeinerte Erfassung der Lagerhaltungskosten wird in fünften Abschnitt eingegangen.

Bei einem Fehlmengenkostensatz von π Geldeinheiten pro Mengen- und Zeiteinheit sind die erwarteten, durch die endliche mittlere Intervallänge $E(S)$ dividierten Fehlmengenkosten im ersten Regenerationsintervall

$$\frac{1}{E(S)}E\left[\pi y\int_0^S (L_y(t)-a)^+\mathrm{d}t\right]$$

$$=\pi y[\int x\mathbf{1}(a<x<\infty)F_L(\mathrm{d}x; y)-a(1-F_L(a; y))].$$

Der Fehlmengenterm in eckigen Klammern auf der rechten Seite der Gleichung stimmt formal mit der erwarteten Fehlmenge während der Lieferfrist im (s, q)-Modell der Lagerhaltungstheorie mit Bestellpunkt bzw. Meldemenge s und fester Losgröße q überein, ([9]), S. 96). Die Verteilungsfunktion beschreibt im (s, q)-Modell allerdings die Verteilung der Nachfrage nach dem betrachteten Produkt in der Lieferfrist, während sich in der hier verwendeten Verteilungsfunktion der Warteschlangenlänge neben der Wirkung der Nachfrage auch der in der Einleitung beschriebene Warteschlangeneffekt in der Produktionssituation spiegelt.

Bezeichnet man mit l die Lagerhaltungskosten pro Mengen- und Zeiteinheit, so ergeben sich die pro Zeiteinheit erwarteten Lagerhaltungskosten analog zu dem Fehlmengenkosten als

$$ly[aF_L(a; y)-\int x\mathbf{1}(0\leq x\leq a)F_L(\mathrm{d}x; y)].$$

Der Ausdruck in eckigen Klammern ist die erwartete Anzahl von Losen im Lager pro Zeiteinheit. Aus ihm

ergibt sich durch Multiplikation mit der Losgröße der erwartete Lagerbestand.

Ist $\{A_y(t), t \geq 0\}$ der Erneuerungsprozeß ([5], S. 109) der Ankünfte, so ist $A_y(S)$ die Anzahl der im ersten Regenerationsintervall angekommenen Produktionsaufträge. Im Vormerkfall ist $A_y(S)$ gleichzeitig die Anzahl der in $[0, S)$ nachgefragten, hergestellten und auch abgesetzten Lose. Der Erwartungswert von $A_y(S)$ ist $E(S)\lambda/y$, die erwarteten Rüstkosten pro Zeiteinheit also $R\lambda/y$, wenn R die durch einen Rüstvorgang verursachten Rüstkosten sind. Die Summe aus Lagerhaltungs-, Fehlmengen- und Rüstkosten

$$K(a; y) = R\lambda/y + (\pi + l)y[aF_L(a; y) - \int x\mathbf{1}(0 \leq x \leq a)F_L(\mathrm{d}x; y)] + \pi y(E(L_y) - a) \quad (1)$$

gibt die erwarteten Kosten pro Zeiteinheit in Abhängigkeit von a und y an.

Der Fehlmengenkostensatz wird in der Regel nur schwer zu ermitteln sein. In diesem Fall kann man die Betrachtung der Fehlmengenkosten durch diejenige eines Servicegrades ersetzen. Unter einem α-Servicegrad wird die Wahrscheinlichkeit dafür verstanden, daß keine Fehlmengen entstehen ([11], S. 105). Fehlmengen können im Modell in einem Zeitpunkt t auftreten, wenn $L_y(t) \geq a$ gilt, da in diesem Fall der Lagerbestand verschwindet und eine eventuell auftretende Nachfrage vorgemerkt werden müßte. Als Grundlage für die Berechnung des hier vorgeschlagenen Servicegrades wird die Wahrscheinlichkeit dieses Ereignisses unter der Grenzverteilung verwendet. Ein vorgegebener Servicegrad von $0 < \gamma < 1$ wird also eingehalten, wenn $P(L_y \leq a - 1) = F_L(a - 1; y) \geq \gamma$ ist.

Die entwickelte Lagerzielmengenpolitik weist eine gewisse Ähnlichkeit zum Vorgehen der bereits erwähnten (s, q)-Politik auf. Beide Politiken verwenden eine feste Losgröße. Allerdings gibt es bei der Lagerzielmengenpolitik nicht nur eine, sondern mehrere Meldemengen, die zur Auslösung eines Fertigungsauftrages führen, nämlich $y(a-1)$, $y(a-2)$ usw. Daher kann bei der Lagerzielmengenpolitik wegen des Warteschlangeneffektes der – insbesondere bei mehreren Sorten interessante – Fall eintreten, daß mehrere Fertigungsaufträge der gleichen Sorte in Bearbeitung sind oder auf diese warten.

3. Die Kostenfunktion bei gegebenen Losgrößen im Mehrsortenfall

Bei der folgenden Übertragung der bisher angestellten Überlegungen auf den Mehrsortenfall werden die im Vorabschnitt verwendeten Symbole analog benutzt, ohne jedesmal neu definiert zu werden. Die Zeit zwischen dem Auftreten der $(j-1)$-ten und j-ten Nachfrage nach einer Produkteinheit der Sorte i wird beispielsweise mit C_{ij} bezeichnet. Es wird o.E. nur der Fall zweier Produkte betrachtet.

Ein gravierender Unterschied gegenüber dem Einsortenfall besteht darin, daß im Mehrsortenfall der betrachtete Warteschlangenlängenprozeß – wie sich noch zeigen wird – kein regenerativer Prozeß mehr ist. Darum sind zusätzliche Überlegungen für die Herleitung der Kostengrößen anzustellen. Für die Anwendung ist der Mehrsortenfall der interessantere, da der in der Einleitung angesprochene Warteschlangeneffekt gegenüber der Herstellung nur eines Produktes hier an Bedeutung gewinnt.

Die Nachfrageseite wird im Mehrsortenfall für alle Produkte (Sorten) für feste Losgrößen wie im Einsortenfall durch unabhängige Folgen von unabhängigen und sortenweise der gleichen Verteilung gehorchenden Zwischennachfragezeiten modelliert. Der Absatz eines Loses und das Auslösen eines Produktionsauftrages läßt sich für jedes der Produkte analog dem Einsortenfall beschreiben, wenn man für die Sorten eine Politik mit Lagerzielmenge a_1 (resp. a_2) betreibt und zwischen der Nachfrage nach den beiden Produkten kein Zusammenhang besteht. Man erhält wie im Fall einer Sorte voneinander unabhängige und identisch verteilte Zwischennachfragezeiten der Lose $A_{i1}, A_{i2}, A_{i3}, \ldots$, für die Sorten $i = 1, 2$.

Ist ein Los einer beliebigen Sorte vollständig abgesetzt, reiht es sich in die Warteschlange vor der Anlage ein; es erhöht die von der entsprechenden Sorte zu berücksichtigende Produktionsmenge um ein Los. Da beide Sorten die gleiche Anlage nutzen, besteht diese Warteschlange als aus „Kunden" zweier verschiedener Typen – Losen der Sorten Eins und Zwei – mit verschiedenen, sortenabhängigen Bearbeitungsdauern. Um in dieser komplizierten Situation dennoch Kostengrößen ermitteln zu können, sind gegenüber dem Einsortenfall zusätzliche Approximationsschritte notwendig. Es wird insofern ein gröberes, die Gegebenheiten weniger genau wiedergebendes Modell betrachtet.

Zunächst wird von den verschiedenen Typen der Kunden (Sorten) abstrahiert und so getan, als kämen nur Lose einer („typischen") Sorte vor der Anlage an. Die verschiedenen Arten von Kunden werden also zu einem einzigen Typ aggregiert. Diese Abstraktion von der Produktsorte ist beim Ankunfts- und beim Bedienvorgang zu berücksichtigen. Die Bediendauern des aggregierten Modells werden als unabhängige, identisch verteilte Zufallvariable $B_1, B_2, \ldots$ gewählt, die mit Wahrscheinlichkeit p_1 die Bearbeitungsdauer eines Loses der Sorte 1, mit der Gegenwahrscheinlichkeit $p_2 = 1 - p_1$ eine solche eines Loses der Sorte 2 als Ausprägung annehmen. $p_i \in (0, 1)$ ist der auf Dauer im Mittel erwartete Anteil von ankommenden Kunden die Art $i \in \{1, 2\}$ in der ursprünglichen Zwei-Sorten-Situation mit gegebenen Losgrößen y_1 bzw. y_2. Die Produktionsdauer eines Loses der Sorte $i \in \{1, 2\}$ setzt sich aus den r_i Zeiteinheiten für den Rüstvorgang, und den y_i/v_i Zeiteinheiten für den eigentlichen Bearbeitungsvorgang zusammen. Im aggregierten Modell nimmt also z. B. die erste Bediendauer B_1 mit Wahrscheinlichkeit p_i den Wert $r_i + y_i/v_i$ an $(i = 1, 2)$. Auf diese Weise bildet das Modell den langfristig erwarteten Anteil der sortenspezifischen Bediendauern der einzelnen Sorten an allen Bediendauern richtig ab. Die erwartete Bediendauer ist

$$E(B_1) = p_1(r_1 + y_1/v_1) + p_2(r_2 + y_2/v_2)$$

und ihre Varianz

$$\mathrm{Var}(B_1) = p_1 p_2 \left(r_1 + \frac{y_1}{v_1} - r_2 - \frac{y_2}{v_2} \right)^2.$$

Bezeichnet man mit $\{N_i(t), t \geq 0\}$ den Erneuerungsprozeß der Ankünfte von Sorte i, so sind von beiden Sorten $N(t) := N_1(t) + N_2(t)$ Kunden bis zum Zeitpunkt t angekommen. $\mathbf{N} := \{N(t), t \geq 0\}$ ist also der Zählprozeß der Ankünfte „typischer" Kunden im aggregierten Warteschlangenmodell. Aus dem Starken Gesetz der Großen Zahlen für Erneuerungsprozesse (vgl. [5], S. 115) folgt, daß die Ankunftsrate von **N** gleich $\nu_1 + \nu_2$ ist, wenn $\nu_i = 1/E(A_{i1})$ der Ankunftsrate von Kunden der Sorte i entspricht. ν_i hängt dabei von der Losgröße der Sorte i ab. Ist nämlich λ_i die mittlere Nachfrage nach den Produkteinheiten der Sorte i pro Zeiteinheit, so werden pro Zeiteinheit im Mittel $\nu_i = \lambda_i / y_i$ Lose bei einer Losgröße von y_i abgesetzt, kommen also vor der Bedienstation des Warteschlangensystems an. Der langfristige mittlere Anteil von Sorte i an allen ankommenden Kunden p_i ist entsprechend $\nu_i/(\nu_1 + \nu_2)$, man kann p_i also aus den Ankunftsraten der beiden Ankunftsprozesse ermitteln.

Das Modell mit „typischen" Kunden besteht aus zwei unabhängigen Ankunftsströmen und einem Schalter (der Anlage) mit Bediendauern, die sich gemäß der Wahrscheinlichkeiten p_1 und p_2 aus den Bearbeitungszeiten der – in diesem Modell zunächst nicht unterscheidbaren – Lose der einzelnen Sorten ergeben. $\mathbf{L} := \{L(t), t \geq 0\}$ bezeichne den Warteschlangenlängenprozeß dieses Systems. Im allgemeinen ist **N** kein Erneuerungsprozeß ([3], S. 293) und **L** kein regenerativer Prozeß mehr ([12], S. 127), so daß im weiteren einige im Einsortenfall nicht benötigte Annahmen, z. B. über die Grenzverteilung von **L**, getroffen werden müssen. Eine Ausnahme ist der Fall, daß die $\{N_i(t), t \geq 0\}$, $i = 1, 2$, unabhängige Poissonprozesse sind. Dann ist **N** ebenfalls ein Poissonprozeß und damit **L** regenerativ ([3], Seite 293).

Um kostenminimale Lagerzielmengen a_i, $i = 1, 2$, herleiten zu können, müssen sortenbezogene Lagerbestands- und Fehlmengengrößen erfaßt werden. Daher müssen die vor der Anlage in einem Zeitpunkt t auf ihre Abwicklung wartenden Produktionsaufträge des aggregierten Modells den beiden Sorten zugeordnet werden. Dieser Disaggregationsvorgang wird als Markierungsvorgang modelliert (vgl. [10], S. 261 ff.): Kommt ein „typischer" Kunde an, so wird er – im Einklang mit der Konstruktion der Bediendauern des aggregierten Modells – mit Wahrscheinlichkeit p_i als Kunde vom Typ i gekennzeichnet. Die Markierungsexperimente sind Bernoulli-Experimente und sollen unabhängig voneinander und von den Zwischenankunftszeiten gestaltet werden. Ergebnis dieses Zuordnungsvorganges ist dann die in t wartende Anzahl von Produktionsaufträgen für Sorte i, $L_i(t)$. Kommen ν $(= \nu_1 + \nu_2)$ Kunden vom aggregierten Typ im Mittel pro Zeiteinheit an, so ist die Ankunftsrate des durch einen solchen Markierungsmechanismus erzeugten Ankunftsstromes von Kunden des Typs i (in Übereinstimmung mit der Ankunftsrate der Ausgangssituation) ν_i. Während die Reihenfolge, in der die Kunden im Markierungsmodell bedient werden, von der Ankunftsreihenfolge in der Ausgangssituation abweicht (vgl. hierzu auch die Überlegungen am Ende dieses Abschnittes), werden die Anteile der Kunden verschiedenen Typs im besprochenen Sinne richtig abgebildet. Speziell gilt dies für den oben erwähnten Fall, daß **N** ein Poissonprozeß ist. Hier bildet nämlich die Anzahl angekommener und als Sorte i markierter Kunden wieder einen Poissonprozeß mit der entsprechenden Rate ([5], S. 111).

Befinden sich n Kunden im Zeitpunkt t im System $(L(t) = n)$, so sind n unabhängige Bernoulli-Experimente mit Erfolgswahrscheinlichkeit p_1 für die Markierung als Kunde vom Typ 1 durchzuführen. Die (bedingte) Wahrscheinlichkeit dafür, daß in t etwa $j \in \mathbb{N}_0$ Kunden vom Typ 1 warten $(L_1(t) = j)$, ist

$$P(L_1(t) = j \mid L(t) = n) = \begin{cases} \binom{n}{j} p_1^j p_2^{n-j} =: b(j; n, p_1) & j \in \{0, 1, 2, \ldots, n\} \\ 0 & \text{sonst,} \end{cases}$$

und die Wahrscheinlichkeit für die Anwesenheit von j Sorte-1-Kunden im System im Zeitpunkt t ist:

$$P(L_1(t) = j) = \sum_{n=j}^{\infty} b(j; n, p_1) P(L(t) = n). \tag{2}$$

Der Erwartungswert der Anzahl von Kunden des Typs 1 im System im Zeitpunkt t ist, da $L(t)$ einen endlichen Erwartungswert hat, $E(L_1(t)) = p_1 E(L(t))$ (analog für Sorte 2). Mit $\eta = (y_1, y_2)$ gilt für die Verteilungsfunktion $F_1(\cdot; t, \eta)$ von $L_1(t)$ für natürliche Zahlen x mit (2)

$$F_1(x; t, \eta) = F_L(x; t, \eta) + \sum_{n>x} P(L(t) = n) \sum_{k=0}^{x} b(k; n, p_1), \tag{3}$$

wobei $F_L(\cdot; t, \eta)$ die Verteilungsfunktion der Warteschlange $L(t)$ im aggregierten, die Sorten nicht unterscheidenden Modell bei den gegebenen Losgrößen der beiden Sorten sei. Kennt man die beiden Ankunftsraten und die Verteilungsfunktion der nach Sorten nicht differenzierten Warteschlangenlänge der „typischen" Kunden, kann man also diejenigen der sortenweisen Warteschlangenlängen daraus ermitteln.

Der Ausgangspunkt für die Herleitung der Kostenfunktion ist im Mehrsortenfall verschieden von demjenigen des Abschnittes 2. Da **L** im allgemeinen kein regenerativer Prozeß ist, werden nicht mehr pro Zeiteinheit erwartete Kosten im ersten Regenerationsintervall, sondern erwartete kumulierte Kosten, wiederum pro Zeiteinheit, betrachtet. Ist a_1 das für Sorte 1 gewählte Lagerniveau, so ist die Fehlmenge in t analog zum Einsortenfall, wenn man nur komplett nachgefragte Lose erfaßt, $y_1(L_1(t) - a_1)^+$. Die bis zu einem Zeitpunkt T kumulierten Fehlmengenkosten betragen dann

$$\pi_1 y_1 \int_0^T (L_1(t) - a_1)^+ \mathrm{d}t.$$

Den auf T Zeiteinheiten bezogenen Erwartungswert dieser Kosten, also die zeitdurchschnittlichen erwarteten Fehlmengenkosten,

$$\pi_1 y_1 \frac{1}{T} E\left(\int_0^T (L_1(t) - a_1)^+ \mathrm{d}t \right)$$

kann man nun mit dem Satz von Fubini bei Beachtung von (3) zu

$$\begin{aligned}\frac{1}{T}\int_0^T E(L_1(t)-a_1)^-\mathrm{d}t &= p_1\frac{1}{T}\int_0^T E(L(t))\mathrm{d}t - a_1\\ &+ a_1\frac{1}{T}\int_0^T P(L(t)\le a_1)\mathrm{d}t\\ &+ a_1\sum_{n>a_1} F_{BV(n,p_1)}(a_1)\frac{1}{T}\int_0^T P(L(t)=n)\mathrm{d}t \qquad (4)\\ &- \sum_{x=0}^{a_1} x\sum_{n\ge x} b(x;n,p_1)\frac{1}{T}\int_0^T P(L(t)=n)\mathrm{d}t\end{aligned}$$

umformen, wobei die Konstante $\pi_1 y_1$ vernachlässigt wurde und $F_{BV(n,p_1)}(x)=\sum_{k=0}^{x} b(k;n,p_1)$ die Verteilungsfunktion der Binominalverteilung ist.

Um die langfristigen zeitdurchschnittlichen erwarteten Kosten, also den Grenzwert von (4) für über alle Grenzen wachsendes T, auf möglichst einfache Weise herleiten zu können, soll – wie angekündigt – die Gültigkeit folgender Annahmen vorausgesetzt werden: **L** hat eine Grenzverteilung, d.h. für alle natürlichen n gilt mit $q_n(t):=P(L(t)=n)$, daß $\lim_{t\to\infty} q_n(t)=q_n\ge 0$, wobei die q_n zu 1 aufsummieren. q_n steht dabei für die Wahrscheinlichkeit $P(L=n)$. Die Verteilungsfunktionen von $L(t)$ mögen im Doppelfolgensinne ([4], S. 251) gegen 1 konvergieren: $\lim_{t\to\infty}\lim_{n\to\infty} F_L(n;t,\eta)=1$, d.h. für alle $\epsilon>0$ gibt es ein $p_\epsilon>0$, so daß für alle $t, n>p_\epsilon$: $|F_L(n;t,\eta)-1|<\epsilon$. Diese Annahme dient dazu, die Vertauschbarkeit von Grenzübergang und Reihenbildung zu sichern und ist daher keine unverzichtbare Forderung. Schließlich möge die Grenzverteilung von **L** einen endlichen Erwartungswert $E(L)$ haben und $\lim_{t\to\infty} E(L(t))=E(L)$.

Die Gültigkeit der Annahmen ist in der Praxis schwer zu prüfen. Es ist daher unter Umständen hilfreich zu wissen, daß die beiden letzten erfüllt sind, falls es eine (feste) maximale Anzahl M von Aufträgen gibt, die auf ihre Bearbeitung warten, so daß $L(t)$ zu jedem Zeitpunkt fast sicher nur Werte kleiner oder gleich M annimmt. Gemeint ist dabei mit M keine obere Grenze für die Anzahl der Produktionsaufträge in dem Sinne, daß ein vor der Anlage ankommender $M+1$-ter Auftrag zurückgewiesen würde. Gedacht ist vielmehr an eine Situation, in der die Nachfrage (Zwischennachfragezeiten) und die Produktion (Bediendauern) solche Form (etwa solche Verteilungen) haben bzw. M so hoch angesetzt wird, daß mit Wahrscheinlichkeit 1 ein M übersteigender Auftragsbestand auch ohne solchen Eingriff nicht erreicht wird. Wegen der Wirkung ökonomischer Mechnismen (etwa des Abwanderns von Nachfrage zu anderen Herstellern eines vergleichbaren Produktes bei langen Lieferzeiten) dürfte ein Modell, das den Auftragsbestand auf Werte unter einer möglicherweise sehr großen Zahl M beschränkt, nichts an praktischer Anwendbarkeit einbüßen.

Unter den genannten Annahmen ist für den ersten Term der rechten Seite von (4)

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T E(L(t))\mathrm{d}t=\sum_{n=0}^{\infty} n\lim_{T\to\infty}\frac{1}{T}\int_0^T q_n(t)\mathrm{d}t=E(L),$$

denn wegen der Existenz der Grenzverteilung geht für alle n auch

$$\frac{1}{T}\int_0^T q_n(t)\mathrm{d}t$$

mit wachsendem T gegen q_n.

Entsprechend argumentiert man auch für die anderen Terme, so daß man die Konvergenz von (4) gegen

$$p_1E(L)-a_1+a_1F_1(a_1;\eta)-\int x\mathbf{1}(0\le x\le a_1)F_1(\mathrm{d}x;\eta), \qquad (5)$$

für wachsendes T erhält. In (5) ist $F_1(\cdot;\eta)$ der Grenzwert von (3) für wachsendes t, also die Verteilungsfunktion der Grenzverteilung der Anzahl von Sorte 1 wartender Kunden. (5) ist die zeitdurchschnittliche erwartete Anzahl von Fehllosen der Sorte 1, durch Multipikation mit dem Fehlmengenkostensatz und der Losgröße der Sorte 1 erhält man daraus die zeitdurchschnittlichen erwarteten Fehlmengenkosten.

Analog kann man für die Lagerhaltungs- und die Rüstkosten vorgehen, wobei die Rüstkosten unabhängig vom angestrebten Lagerbestandsniveau a_1 sind. Das Kriterium für die Wahl des Lagerbestandsniveaus lautet dann (vgl. (1)):

$$\begin{aligned}R_1\lambda_1/y_1 &+ (\pi_1+l_1)y_1[a_1F_1(a_1;\eta)\\ &-\int x\mathbf{1}(0\le x\le a_1)F_1(\mathrm{d}x;\eta)]+\pi_1y_1(E(L_1)-a_1).\end{aligned}$$

Durch die Addition der entsprechenden Kosten der Sorte 2 erhält man die bezüglich der beiden Lagerniveaus zu minimierenden, auf Dauer pro Zeiteinheit erwarteten Kosten

$$\begin{aligned}K(a_1,a_2;\eta)=\sum_{i=1}^{2}\{&R_i\lambda_i/y_i+(\pi_i+l_i)y_i[a_iF_i(a_i;\eta)\\ &-\int x\mathbf{1}(0\le x\le a_i)F_i(\mathrm{d}x;\eta)]\\ &+\pi_iy_i(E(L_i)-a_i)\}.\end{aligned} \qquad (6)$$

Die Kosten in (6) stellen in gewissem Sinne eine obere Schranke für die eintretenden Kosten dar. Denn in dem ursprünglichen Modell mit sortenspezifischen Ankünften wird die Bearbeitungsreihenfolge der Lose durch die Ankunftsreihenfolge festgelegt, die im aggregierten Modell durch das Zufallsexperiment der Sortenwahl abgebildet wird. Die betriebliche Planung wird aber auf eine Festlegung der Bearbeitungsreihenfolge der Lose nicht verzichten wollen, um etwa die Dringlichkeit der Lagerzuführung einer Sorte (bei geringem Lagerbestand), relativ hohe Lagerbestände einzelner Sorten oder von der Fertigungsreihenfolge abhängige Rüstkosten zu berücksichtigen. Wird diese Planung sinnvoll durchgeführt, sollten die Kosten niedriger als in (6) anfallen.

4. Approximation der Kostenfunktion im Mehrsortenfall

Im allgemeinen werden die zur Berechnung der Kostenfunktion benötigten Verteilungsfunktionen und Erwartungswerte nicht bekannt sein, so daß die Notwendigkeit besteht, zu geeigneten Näherungsgrößen zu gelangen. Entscheidend für die Approximation ist die von der Losgröße abhängige (gemeinsame) Verkehrsintensität

$$\varrho(y_1, y_2) = \frac{r_1\lambda_1}{y_1} + \frac{\lambda_1}{v_1} + \frac{r_2\lambda_2}{y_2} + \frac{\lambda_2}{v_2}. \tag{7}$$

Sie ergibt sich aus dem Quotienten der Ankunftsrate von **N** und der Bedienrate in dem Modell mit „typischen" Kunden. Sowohl für die ursprüngliche 2-Sorten-Situation als auch für das aggregierte Modell mit Markierungsexperiment läßt sich für $\varrho(\eta) < 1$, ohne die eingeführten Annahmen zu nutzen, zeigen, daß $E\left[\int_0^T \mathbf{1}(L(t) > 0)\mathrm{d}t\right]/T$, also der erwartete Zeitanteil, währenddessen die Anlage in Betrieb ist, mit wachsendem T gegen (7) konvergiert (vgl. Anhang). Hierfür hat man lediglich zu unterstellen, daß bei mindestens einer Sorte die maximale Zeitdauer, die zwischen je zwei Nachfragevorgängen vergeht, fast sicher beschränkt ist. Damit läßt sich auch im Mehrsortenfall die Verkehrsintensität als auf Dauer sich einstellender erwarteter Zeitanteil interpretieren, der für Rüst- und Herstellvorgänge benötigt wird, also als Auslastungsgrad. Für die Approximation soll unterstellt werden, daß der Auslastungsgrad nur wenig kleiner als 1 ist; denn im Falle niedriger Auslastung kann man der Tendenz nach die Sorten isoliert betrachten, jeweils eine nicht beschränkte Produktionsgeschwindigkeit annehmen und daher den in der Einleitung angesprochenen (einfacheren) Ansatz mit diskreter Zeit heranziehen.

Losgrößen mit niedriger Kapazitätsauslastung können ferner nicht sinnvoll sein, da Unternehmen bestrebt sein werden, ihre Fertigungskapazität auf möglichst hohem Niveau auszulasten, weil die Aufrechterhaltung der Kapazität Kosten verursacht. Wird die Kapazität auf Dauer nur zu einem geringen Teil genutzt, so wird man sie folglich reduzieren. Da die Kapazitätsanpassung hier ausgeblendet bleiben soll, wird unterstellt, daß die Kapazität schon sinnvoll gewählt ist, d. h. nur noch hohe Auslastungsgrade relevant sind. Unter hohen Auslastungsgraden werden dabei solche verstanden, die oberhalb einer vorgegebenen, unternehmensindividuellen unteren Auslastungsschranke von ϱ_u liegen, $\lambda_1/v_1 + \lambda_2/v_2 < \varrho_u < 1$.

Zu hohen Auslastungsgraden gehören gemäß Formel (7) Losgrößen, die im Vergleich zu solchen bei einer niedrigen Auslastung klein sind. Diese relativ kleinen Lose haben den Vorteil einer relativ große Auflagehäufigkeit. Der Beginn jedes Rüstvorganges ist nämlich ein natürlicher Eingriffszeitpunkt in den betrieblichen Herstellungsprozeß. Eine Erhöhung ihrer Anzahl bietet die Chance zur, gerade bei Unsicherheit notwendigen, Anpassung des Fertigungsablaufes an die jeweils aktuelle Datensituation.

Es gibt auch Gründe, die unter Umständen eine Erhöhung der Losgrößen wünschenswert erscheinen lassen. Etwa den Effekt der Rüstkostendegression oder den der Substitution von Rüstzeiten durch produktive Maschinennutzungszeiten (sofern die Kapazität der betrachteten Maschine knapp ist). Diese Aspekte werden im Modell aber explizit – durch Berücksichtigung von Rüstzeiten und Rüstkosten einerseits und der die Kapazitätsknappheit widerspiegelnden Bedingung $\varrho(\eta) < 1$ andererseits – erfaßt.

Bei hohem Auslastungsgrad kann man den Warteschlangenlängenprozeß der „typischen" Kunden **L** durch eine regulierte Brownsche Bewegung $\mathbf{W} = \{W(t), t \geq 0\}$ approximieren ([7], S. 360 und 363f.), d. h. im Kriterium (4) im vorstehenden Abschnitt $E(L(t))$ und $F_L(\cdot; t, \eta)$ durch die entsprechenden Größen einer Brownschen Bewegung $E(W(t))$ und $F_W(\cdot; t, n)$ ersetzen, ferner $P(L(t) = n)$ in geeigneter Weise – z. B. durch $F_W(n; t, \eta) - F_W(n-1; t, \eta)$ – annähern ($W(t)$ ist ja eine stetige Zufallsvariable; beachte für $n = 0$ auch die letzte Bemerkung des Anhanges).

Anschließend kann man den Limes der so modifizierten Formel (4) untersuchen. Der Grenzwert, gegen den der Erwartungswert unter der Näherung für wachsendes t konvergiert, ist

$$E(W) := \lim_{t \to \infty} E(W(t)) = \frac{1}{2(1 - \varrho(y_1, y_2))}\left[\tau_B^2 + \varrho(y_1, y_2)\tau_A^2\right]$$

([5], S. 479), wobei τ_A eine Art Variationskoeffizient der Zwischenankunftszeiten (vgl. unten), τ_B den Variationskoeffizienten der Bediendauern im aggregierten Modell bezeichne. Sowohl τ_A als auch τ_B hängen von den Losgrößen y_1 und y_2 ab, auch wenn dies aus Gründen der Übersichtlichkeit im verwendeten Symbol nicht deutlich wird:

$$\tau_A^2 = p_1^3 \operatorname{Var}(A_{11}) + p_2^3 \operatorname{Var}(A_{21}) \Big/ \left(\frac{1}{v_1 + v_2}\right)^2$$

$$= \left(\frac{\lambda_1}{y_1} + \frac{\lambda_2}{y_2}\right)^{-1}\left[\frac{\lambda_1^3 \operatorname{Var}(C_{11})}{y_1^2} + \frac{\lambda_2^3 \operatorname{Var}(C_{21})}{y_2^2}\right]$$

und

$$\tau_B^2 = p_1 p_2 \left(r_1 + \frac{y_1}{v_1} - r_2 - \frac{y_2}{v_2}\right)^2 \Big/ \left[p_1\left(r_1 + \frac{y_1}{v_1}\right) + p_2\left(r_2 + \frac{y_2}{v_2}\right)\right]^2.$$

Mit $E(W)$ kann man die Grenzverteilungsfunktion als

$$F_W(x; \eta) = 1 - \exp\left\{-\frac{x}{E(W)}\right\}$$

schreiben. Hieraus lassen sich die Grenzwerte der in (4) auftretenden Terme errechnen. Geht man für die Lager-

haltungs- und Rüstkosten ebenso vor, ergibt sich wieder die Kostenfunktion (6), wobei Erwartungswerte und Verteilungsfunktionen durch ihre eben entwickelten Approximationen zu ersetzen sind.

Alternativ gelangt man zu demselben zu optimierenden Zielkriterium, nämlich den hergeleiteten Näherungsgrößen eingesetzt in (6), auch durch einem weiteren Ansatz. Zentrales Problem des Warteschlangensystems mit zwei Ankunfsquellen ist, daß es in der Regel kein regeneratives System mehr ist. Whitt [12] konstruiert daher zu solchen nicht notwendig regenerativen Systemen Ersatzsysteme, die diese Struktureigenschaft aufweisen. Sie werden als Warteschlangensysteme mit nur einem Erneuerungsprozeß als Ankunfsstrom angelegt, wobei die Verteilung der Zwischenankunfszeiten so gewählt wird, daß die Ein-Quellen-Systeme eine möglichst gute Annäherung an die Ausgangssysteme ergeben. Für die hier vorliegende Situation schlägt Whitt mit seiner asymptotischen Methode ([12], S. 131f.) vor, als Erwartungswert der neuen Zwischenankunftszeiten $1/\nu = 1/(\nu_1 + \nu_2)$ und als ihre Varianz $\nu_1^3 \operatorname{Var}(A_{11}) + \nu_2^3 \operatorname{Var}(A_{21})$ zu wählen. Unter Ausnutzung der Regenerativität des so konstruierten approximierenden Warteschlangensystems kann man das Kostenkriterium (6) analog zum Einsortenfall in Abschn. 2 herleiten.

Bei der Übertragung von Whitts Vorgehen auf die hier vorliegende Situation, spielte bislang die Form der Verteilung der Zwischenankunfszeiten im Ersatzsystem keine Rolle. Das ändert sich, wenn man den zu bestimmten Lagerzielmengen und Losgrößen gehörenden Wert der Kostenfunktion für eine gegebene Datenkonstellation berechnen will. Denn selbst bei Verteilungskenntnis hat man im allgemeinen Probleme, daraus die Verteilungsfunktion der Grenzverteilung des Warteschlangenlängenprozesses zu bestimmen, die man für die anzustellenden Kalkulationen benötigt. Ist aber wieder die Auslastung des approximierenden, regenerativen Systems hoch, kann man auf die schon verwendete Brownsche Approximation zurückgreifen, also z. B. den unbekannten Erwartungswert der Grenzverteilung durch $E(W)$ annähern (vgl. [5], S. 482).

Die Näherungsgrößen entsprechen den oben angeführten, wobei nun auch die obige Bezeichnung von τ_A als Variationskoeffizient klar wird: τ_A ist der Variationskoeffizient der Zwischenankunftszeiten in dem nach Whitt gewählten Ersatzsystem.

5. Bestimmung optimaler Lagerzielmengen und Losgrößen

Für die verschiedenen Optimierungsüberlegungen soll unterstellt werden, daß L, wie in Abschn. 4 dargestellt, durch eine geeignete regulierte Brownsche Bewegung angenähert wird. Es werden keine Fehlmengenkosten betrachtet; stattdessen werden die interessierenden Größen so bestimmt, daß die Summe aus Lagerhaltungs- und Rüstkosten minimiert wird unter der Nebenbedingung, daß ein vorgegebener, sortenspezifischer Servicegrad von γ_1 bzw. γ_2, $0 < \gamma_1, \gamma_2 < 1$, gilt. Analog zum Einsortenfall ist die Einhaltung des Servicegrades z.B. für Sorte 1 gewährleistet, wenn $P(L_1 \leq a_1 - 1) = F_1(a_1 - 1; \eta) \geq \gamma_1$ ist.

Bei gegebenen Losgrößen $\eta = (y_1, y_2)$ setzen sich die bei der Wahl der Lagerzielmengen zu betrachtenden Kosten aus zwei unabhängigen, additiv verknüpften Lagerhaltungs- und Rüstkostentermen für die beiden Sorten zusammen. Da in diesem Fall auch die Servicegrade der Sorten nicht voneinander abhängen, kann man jede Sorte isoliert untersuchen. Nun ist die Summe aus Rüst- und Lagerhaltungskosten etwa der Sorte 1 monoton nichtfallend in der Lagerzielmenge, so daß das kleinste ganzzahlige a_1, das die Servicegradbedingung einhält, diese Kosten minimiert (analog für Sorte 2). Die optimalen Lagerzielmengen lassen sich daher auf numerischem Wege durch Auswertung der Servicegradrestriktion bestimmen.

Bei der simultanen Optimierung von Lagerzielmengen und Losgrößen muß man weitergehende Überlegungen anstellen, da im bisher entwickelten Modell die für die Loswahl wichtigen Lagerhaltungskosten nur grob erfaßt werden. Der Lagerbestand z. B. der Sorte 1 wächst nach den bisherigen Überlegungen erst, wenn ein Los dieser Sorte vollständig bearbeitet ist. Während der Bearbeitungsdauer ist aber im Mittel ein halbes Los von Sorte 1 bei der Anlage vorhanden. Diese Lagermenge wird im bisherigen Modell nicht erfaßt (vgl. auch [8]).

Mit Hilfe des Starken Gesetzes der Großen Zahlen für Renewal-Reward-Prozesse ([5], S. 173) läßt sich zeigen, daß der erwartete Zeitanteil, währenddessen etwa Sorte 1 gefertigt wird, im Markierungsmodell mit wachsendem T gegen $q_1(\eta)\varrho(\eta)$ mit $q_1(\eta) := p_1 \; (r_1 + y_1/v_1)/E(B_1)$ geht, also gegen den erwarteten Zeitanteil, in dem überhaupt gefertigt wird, multipliziert mit dem Anteil der erwarteten Bediendauer von Sorte 1 an der erwarteten Gesamtbediendauer. Die entsprechende, zusätzlich zu berücksichtigende Lagermenge für Sorte 1 damit $q_1(\eta)\varrho(\eta)y_1/2$.

Andererseits wurde in Abschn. 2 dargestellt, daß der Lagerbestand, solange er von Null verschieden ist, im Modell überschätzt wird (vgl. auch Abb. 1 in Abschn. 2). Denn während dieser Zeit befindet sich ein Los gerade im Absatzprozeß, ist also nur zu einem Teil tatsächlich im Lager vorhanden. Das Modell überschätzt den Lagerbestand während dieser Zeit im Mittel z. B. für Sorte 1 um $(y_1 - 1)/2$ Mengeneinheiten, da es so tut, als ob das Los vollständig wäre. Der Zeitanteil, währenddessen der Lagerbestand nicht verschwindet, läßt sich durch $P(L_1 < a_1) = F_1(a_1 - 1; \eta)$ approximieren, so daß der Lagerbestandsterm zur genaueren Erfassung der Lagerhaltungskosten um $(y_1 - 1)F_1(a_1 - 1; \eta)/2$ zu bereinigen ist. Der modifizierte, pro Zeiteinheit auf Dauer erwartete Lagerbestand von Sorte 1 ist dann

$$l_1(a_1, \eta) = y_1[a_1 F_1(a_1; \eta) - \int x \mathbf{1}(0 \leq x \leq a_1) F_1(\mathrm{d}x; \eta) + \frac{1}{2} q_1(\eta)\varrho(\eta)] - \frac{y_1 - 1}{2} F_1(a_1 - 1; \eta).$$

Die Summe aus Rüst- und modifizierten Lagerhaltungskosten der Sorte 1, $\lambda_1 R_1/y_1 + l_1(a_1, \eta)$, ist für jedes gegebene Losgrößenpaar η monoton nichtfallend in $a_1 \in \mathbb{R}_+$. Damit ist die kostenminimale Lagerzielmenge bei beliebig gegebenem η, $a_1^*(\eta)$, die kleinste den vorgegebenen Servicegrad γ_1 einhaltende Menge. Analoge Überlegungen gelten für Sorte 2.

Im Anwendungsfall wird man immer eine obere Schranke für die Losgrößen der Sorten angeben können, etwa die Jahresfertigungskapazität. Dann sind nur für endlich viele Losgrößenpaare solche kostenminimalen Lagerzielmengen zu bestimmen. Anschließend ist für die so gewonnenen, endlich vielen Kombinationen aus Losgrößen und Lagerzielmengen $(a_1^*(y_1,y_2), a_2^*(y_1,y_2), y_1, y_2)$ ein Kostenvergleich durchzuführen, um die optimale Kombination auszuwählen.

Die tatsächliche numerische Ermittlung der optimalen Losgrößen und Lagerzielmengen ist allerdings nicht anspruchslos, da z. B. die in den Kosten- und Servicegradtermen vorkommenden Verteilungsfunktionen wie beschrieben approximiert und ausgewertet werden müssen. Hier liegt – neben der numerischen Erprobung – ein Feld weiterer Untersuchungen im Zusammenhang mit dem in dieser Arbeit vorgestellten Ansatz. Das gilt auch für Überlegungen bezüglich der Einbettung des Modells in praktische Planungsvorgänge und die daraus erwachsende Notwendigkeit eines revidierenden Einsatzes der vorgeschlagenen Verfahren.

Anhang

Sei $(\Omega, \mathscr{L}, P)$ der zugrunde liegende Wahrscheinlichkeitsraum. Wir treffen folgende *Annahme (A):* Die Zeit, die zwischen je zwei Nachfragevorgängen vergeht, ist bei mindestens einer Sorte fast sicher (f.s.) kürzer als δ Zeiteinheiten.

Bei gegebenen Losgrößen beläuft sich die gesamte Bearbeitungszeit der bis zum Zeitpunkt t erteilten Fertigungsaufträge auf $G(t) := (r_1 + y_1/v_1)N_1(t) + (r_2 + y_2/v_2)$ $N_2(t)$ Zeiteinheiten im ursprünglichen 2-Sorten-Modell (und auf $G(t) := \sum_{i=1}^{N(t)} [(r_1 + y_1/v_1)\mathbf{1}(X_i = 1) + (r_2 + y_2/v_2)$ $\mathbf{1}(X_i = 2)]$ im Markierungsmodell; dabei seien $X_1, X_2, \ldots$ identisch verteilte und stochastisch unabhängige Markierungsvariable mit $P(X_1 = 1) = p_1$ und $P(X_1 = 2) = p_2 = 1 - p_1$). Die tatsächliche Betriebszeit der Anlage bis t ist $I(t) = \int_0^t \mathbf{1}(L(s) > 0)\mathrm{d}s$, denn die Anlage ist in Betrieb, solange Kunden im System sind. In beiden Fällen ist $I(t) \le G(t)$ für alle $t \ge 0$ und nach dem Starken Gesetz der Großen Zahlen für Erneuerungsprozesse konvergiert $G(t)/t$ mit t f.s. gegen $(r_1 + y_1/v_1)p_1\nu + (r_2 + y_2/v_2)p_2\nu = \varrho(\eta) =: \varrho$. Also ist f.s. $\limsup_{t\to\infty} \frac{1}{t} I(t) \le \varrho$.

Gilt (A) und $\varrho < 1$, so gibt es für fast alle ω eine (strikt wachsende) Folge von „Leerzeitpunkten" $\{S_n(\omega), n \in \mathbb{N}\}$ mit $L(\omega, S_n(\omega)) = 0$. Denn gibt es eine Menge $M \in \mathscr{L}, P(M) > 0$, so daß für jedes ω aus M schließlich $L(\omega, t)$ strikt positiv ist, so geht $I(\omega, t)/t$ auf M gegen 1 – ein Widerspruch zur obigen Aussage über den Limes Superior.

Behauptung: *Gilt (A) und $\varrho < 1$, so folgt*

$$(1)\quad \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{1}(L(s) > 0)\mathrm{d}s = \varrho \text{ f.s.};$$

$$(2)\quad \lim_{t\to\infty} \frac{1}{t} E\left(\int_0^t \mathbf{1}(L(s) > 0)\mathrm{d}s \right) = \varrho.$$

Beweis: Zu **(1):** Die Pfade von **L** liegen f.s. in $D[0,\infty)$ und sind f.s. Sprungfunktionen. Daher kann man spezielle Leerzeitpunktfolgen wählen. Für fast alle ω gilt: Man kann die nichtnegative Zeitachse in linksseitig abgeschlossene, rechtsseitig offene Intervalle zerlegen, über denen $L(\omega, \cdot)$ jeweils konstant ist. Unter ihnen sind insbesondere solche, über denen $L(\omega, \cdot)$ verschwindet. $\{S_n(\omega), n \in \mathbb{N}\}$ sei nun die Folge der linken Randpunkte aller dieser Intervalle (aufsteigend angeordnet).

Für die Zeitpunktfolge $\{t_m, m \in \mathbb{N}\}$ sei $S_{n(m)}(\omega)$ der späteste vor t_m gelegene Leerzeitpunkt, so daß t_m im Intervall $[S_{n(m)}(\omega), S_{n(m)+1}(\omega))$ liegt (bis auf höchstens endlich viele Ausnahmezeitpunkte). In den Leerzeitpunkten ist der Anteil der Betriebs- an der Gesamtzeit $I(\omega, S_{n(m)}(\omega))/S_{n(m)}(\omega) = G(\omega, S_{n(m)}(\omega))/S_{n(m)}(\omega)$. Erhöht man von einem solchen Zeitpunkt ausgehend den Nenner t kontinuierlich, so sinkt dieser Zeitanteil, denn der Nenner beider Quotienten wächst, während der Zähler konstant bleibt, solange keine Ankunft eines neuen Kunden (Loses) stattfindet. Steht Δ für δy, y für die Losgröße derjenigen Sorte, die die Annahme (A) erfüllt, so kommt spätestens Δ Zeiteinheiten nach dem Leerzeitpunkt ein Kunde an. Also ist im Intervall $[S_{n(m)}(\omega) + \Delta, S_{n(m)+1}(\omega))$ $L(\omega, t) > 0$ und der Betriebszeitanteil wächst. Eine untere Schranke für den Betriebszeitanteil im Intervall $[S_{n(m)}(\omega), S_{n(m)}(\omega) + \Delta)$ ist folglich $b_m(\omega) := G(\omega, S_{n(m)}(\omega))/(S_{n(m)}(\omega) + \Delta)$.

$$b_m(\omega) \text{ ist aber gleich } \frac{S_{n(m)}(\omega)}{S_{n(m)}(\omega)+\Delta} \frac{1}{S_{n(m)}(\omega)} G(\omega, S_{n(m)}(\omega)),$$

geht also mit m gegen ϱ. Es folgt: $\liminf_{t\to\infty} \frac{1}{t} I(t) \ge \varrho$ f.s. und damit (1).

Zu **(2):** (2) folgt aus (1) durch die Anwendung des Satzes von der majorisierten Konvergenz. □

Aus (2) folgt $\lim_{t\to\infty} \frac{1}{t} \int_0^t P(L(s) > 0)\mathrm{d}s = \varrho$. Existiert eine Grenzverteilung von **L**, ergibt sich also unter den Voraussetzungen der Behauptung $P(L = 0) = 1 - \varrho$.

Literatur

1. Altiok T (1985) (R, r) Production/inventory systems. Oper Res 37:266–276
2. Gaver DP (1961) Operating characteristics of a simple production, inventory-control model. Oper Res 9:635–649
3. Grimmett GR, Stirzacker DR (1982) Probability and random processes. Oxford University Press, Oxford
4. Heuser H (1980) Lehrbuch der Analysis, Teil 1. Teubner, Stuttgart
5. Heyman DP, Sobel MJ (1982) Stochastic models in operations research, vol I. McGraw-Hill, New York
6. Heyman DP, Sobel MJ (1984) Stochastic models in operations research, vol II. McGraw-Hill, New York

7. Iglehart DL, Whitt W (1970) Multiple channel queues in heavy traffic. II: Sequences, networks, and batches. Adv Appl Probab 2:355–369
8. Jahnke H (1991) Eine Lagerzielmengenpolitik bei losweiser Fertigung und unsicherer Nachfrage im Einsortenfall. Arbeitspapier Nr. 27 des Institutes für Logistik und Transport der Universität Hamburg
9. Kistner K-P, Steven M (1990) Produktionsplanung. Physica-Verlag, Heidelberg
10. Kleinrock L (1975) Queueing systems, vol I: Theory. Wiley, New York
11. Klemm H, Mikut M (1972) Lagerhaltungsmodelle. Verlag Die Wirtschaft, Berlin
12. Whitt W (1982) Approximating a point process by a renewal process, I: Two basic methods. Oper Res 30:125–147
13. Williams TM (1984) Special products and uncertainty in production/inventory systems. Eur J Oper Res 51:46–54
14. Zipkin PH (1986) Models for design and control of stochastic, multi-item batch production systems. Oper Res 34:91–104