A Study of Moment Recursion Models for Tactical Planning of a Job Shop: Literature Survey and Research Opportunities

Chee Chong Teo
Singapore-MIT Alliance, N2-B2C-15, Nanyang Technological University, 50, Nanyang Ave,
Singapore 639798

***Abstract* — The Moment Recursion (MR) models are a class of models for tactical planning of job shops or other processing networks. The MR model can be used to determine or approximate the first two moments of production quantities and queue lengths at each work station of a job shop. Knowledge of these two moments is sufficient to carry out a variety of performance evaluation, optimization and decision-support applications. This paper presents a literature survey of the Moment-Recursion models. Limitations in the existing research and possible research opportunities are also discussed. Based on the research opportunities discussed, we are in the process of building a model that attempts to fill these research gaps.**

***Index Terms*— job shop tactical planning model, literature survey and research opportunities in moment recursion models, moments of production quantities and queue lengths.**

I. Introduction

THIS paper considers a class of models for tactical planning of job shops or other processing networks. This class of models is called the Moment Recursion (MR) models. A job shop is a process structure in which there is a wide variety of jobs and a jumbled work flow through the shop. Due to the large variety of jobs and the diverse processing requirements of each job, there is no distinct workflow through the shop. Specifically, a work station may receive jobs from different stations and jobs completed at the station may be routed to one of the several stations or may leave the shop if completed. Because of the wide variety of jobs and thus a lack of prevailing work flow, production control is difficult and can be very complex. Examples of job shops are the manufacturing of custom-made products, such as machining workshops, where each job has unique requirements from different customers.

A job shop often represents the most complex and generic form of a manufacturing environment. Therefore, the ability to plan a job shop will provide useful insights for production control of such other process structures.

By tactical planning, we imply that we are not concerned with the detailed scheduling issues. Here, we are more interested in identifying the dominant flows in the job shop and subsequently modeling the job shop for tactical planning purposes, such as capacity planning.

We consider a class of models called the MR Models for modeling and analyzing job shops. The defining features of a MR model are:

1) It is a discrete time model, such that work is completed during fixed-length periods, and work arrivals and transfers occur at the start of these periods.
2) Each station of the network produces a quantity that depends on the work-in-process level at that station through some production function.
3) Workflows are Markovian such that processing requirements do not depend on how the work gets to the station.
4) Work arrivals are stochastic and are characterized by finite mean and variance.
5) Recursion equations can be written to describe the relationship between the first two moments of production quantities and queue lengths in periods.

The MR models can be used to determine or approximate the first two moments of production quantities and queue lengths at each work station of a job shop. Knowledge of the first two moments is sufficient to carry out a variety of performance evaluation, optimization and decision-support applications.

For example, we may use MR models to help optimize processing networks of job shops. We may optimize the performance of a job shop (e.g. throughput rate or work-in-process inventory) with a budget constraint for capacity, or minimize the capacity cost with a performance target (throughput rate, work-in-process or lead time).

One of the main capabilities of the MR models is its ability to compute the second moment of production volume, which is one of the key measures of variability in

Manuscript received 3rd November 2003.

C.C. Teo is a PhD student with the Singapore-MIT Alliance (SMA) under the Innovations in Manufacturing Systems and Technology (IMST) program hosted at Nanyang Technological University, Singapore. (email:ps7516907h@ntu.edu.sg).

a manufacturing facility. Variability of production volume is important for several reasons. A larger degree of variability implies a higher level of finished-goods inventory for a make-to-order manufacturer, or a longer lead-time quoted to customers in a make-to-order environment. In addition, variability can upset production planning and schedules. Overtime, use of subcontractors, temporary workers and other expediting services are often associated with high variability.

This paper presents a literature survey of the Moment-Recursion models which is useful for modeling and analyzing job shops. Limitations in the existing research and some possible research opportunities are discussed. It is hoped that such a discussion would provide directions for future work to make MR models more realistic and applicable to real-life manufacturing environment. The remainder of the paper is organized into three sections. Section II gives a review of the concept of a MR model. Section III presents the literature survey of MR models as well as other related research. Section IV discusses the limitations and possible research opportunities in this area.

II. Concept of a Moment-Recursion Model

In this section, we present the basic concept of a MR model. There are a number of variations within this class of models, largely in the control rules and release policies. The different variations will be discussed in the next section. Here, we will only look at the MR models in its most general form. Consider a job shop or other network of processing stations. The shop is modeled in discrete time, such that the stations process some amount of work-in-process during each time period, and transfer the work to other stations or out of the network at the start of the next time period. The workflow is modeled as work hours, rather than as a set of distinct jobs.

Each station i must satisfy the following elementary inventory balance equation,

$$Q_{i,t} = Q_{i,t-1} - P_{i,t-1} + A_{i,t} \qquad (1)$$

where $Q_{i,t}$ is the work-in-process level (both in queue and in service) at the start of period t, $P_{i,t-1}$ is the amount of work processed in period *t-1*, and $A_{i,t}$ is the amount of work that enters station i at the start of period t. We write (1) for all stations in matrix form as

$$\boldsymbol{Q_t} = \boldsymbol{Q_{t-1}} - \boldsymbol{P_{t-1}} + \boldsymbol{A_t} \qquad (2)$$

where $\boldsymbol{Q_t}$ is a vector of work-in-process, $\boldsymbol{P_{t-1}}$ is a vector of production quantities, and $\boldsymbol{A_t}$ is a vector of arrivals.

The following assumptions are made about $\boldsymbol{P_t}$ and $\boldsymbol{A_t}$. $\boldsymbol{P_t}$ is a function of the work-in-process at the start of period t, i.e. $\boldsymbol{P_t} = \boldsymbol{f}(\boldsymbol{Q_t})$, where f is a production function or control rule that relates the production quantities of a station in a period with the work-in-process at the start of the same period. $\boldsymbol{A_t}$ has two components. The first component $\boldsymbol{A_t^I}()$ represents work that is transferred *internally* within the network between stations and is a deterministic function of $\boldsymbol{P_{t-1}}$, i.e. $\boldsymbol{A_t^I}(\boldsymbol{P_{t-1}})$. It corresponds to the concept that completed work at one station in one period triggers work at other stations in the next period. The second component $\boldsymbol{A_t^E}$ consists of work that arrives to stations *externally* from outside the shop. $\boldsymbol{A_t^E}$ can be independent and identically distributed, or can be a function of $\boldsymbol{Q_{t-1}}$ i.e. $\boldsymbol{A_t^E}(\boldsymbol{Q_{t-1}})$ which allows the modeling of “pull” release policies such as constant-inventory job shops.

By substituting the above expressions for $\boldsymbol{P_{t-1}}$ and $\boldsymbol{A_t}$ into (2), we obtain the recursion equation (3) that that forms the basis of the MR models.

$$\boldsymbol{Q_{i,t}} = \boldsymbol{Q_{i,t-1}} - \boldsymbol{P_{i,t-1}}(\boldsymbol{Q_{t-1}}) + \boldsymbol{A_t^I}[\boldsymbol{P_{t-1}}(\boldsymbol{Q_{t-1}})] + \boldsymbol{A_t^E} \qquad (3)$$

If we know the expected value and variance of $\boldsymbol{A_t^I}$, we can express (3) in terms of $\boldsymbol{P_t}$ and $\boldsymbol{P_{t-1}}$, or alternatively $\boldsymbol{Q_t}$ and $\boldsymbol{Q_{t-1}}$. By repeatedly iterating the resulting close-form expressions, and assuming that the job shop has a defined steady state, we can obtain a converging infinite series. We can determine the expected values and the variance of this series to obtain the first two moments of $\boldsymbol{P_t}$ and $\boldsymbol{Q_t}$.

III. Related literature

We devote this section to a literature survey of MR models as well as other related models of processing networks.

A. Moment-Recursion Models

The first paper on the MR models is by Graves in which he developed the Tactical Planning Model (TPM) [1]. The stations in this model use a linear control rule $P_{i,t} = \alpha_i Q_{i,t}$, where α_i is a production smoothing parameter for station i and its inverse $1/\alpha_i$ is the planned lead time. The control rule implies that the production rate at station i is a fixed proportion α_i of its queue length in each time period, and is consistent with the assignment of a planned lead time to each station. The longer the planned lead time, the greater amount of production smoothing the station will be subjected to. This control rule is based on the approximate analytical model for production smoothing developed by Cruickshanks, Drescher and Graves [2]. By using an example of a factory that produces grinding machines, he illustrated how to use the model to evaluate the choice of the planned lead time and also to find a good specification that will result in an acceptable shop behavior.

Parrish presented some extensions to the TPM [3]. First, he proposed a framework for modeling work releases to meet a delivery due date for a finished product. In addition, he showed how to apply the TPM modeling framework to generate two service measures - the probability that demand exceeds inventory and the average number of successive periods in which demand is not met. In addition, he also showed how to adjust the control parameters of the TPM to change these service measures.

Leong modeled the Kanban and other pull systems using the TPM [4]. In a pull system, work is produced at a station whenever there is a downstream inventory shortfall. Here, the linear control rule is $P_{i,t} = \alpha_i (T_i - Q_{i,t})$, where T_i is the target inventory level.

Graves presented a model to provide a rough-cut assessment of both the staffing and component inventory levels of a repair depot [5]. The repair rate of the depot has a production function that resembles the linear control rule of the TPM. He first suggested the piecewise linear production function $P_t = \min[Q_t, K + Q_t / n]$ where K and n are constants. The rationale behind this function is that the total production level cannot exceed the work-in-process (backlog of failed units in this case). Here, the production level is set equal to the sum of a constant term K and a term that is proportional to the work-in-process Q_t/n. However, this function cannot be evaluated directly. Hence, he suggested an approximation $P_t = K + Q_t / n$. Here the values of K and n are set so that the probability of the production quantities P_t exceeding the work-in-process Q_t is small.

Graves presented three extensions to a single station model of the TPM [6]. First, he modeled a station that fails according to a Bernoulli process and the duration of each failure is exactly one period Second, he derived the approximate steady state moments for the TPM with lot-sizing. In this model, work completed by the station is merged into lots of fixed size, and the lots are routed probabilistically. Finally, he presented the mathematical bounds on the behavior of a station with a bounded control rule which depicts the capacity constraint at the station. The control rule is of the form $P_t = \min(\alpha Q_t, M)$, where M is the capacity constraint of the station.

Mihara extended the work of Graves [6] on unreliable single station in TPM when he looked at unreliable multi-station TPM [7]. But similar to Graves' work, the stations also fail according to a Bernoulli process. He also performed simulation studies of a multi-station TPM in which each station i uses bounded control rules of the form $P_t = \min(\alpha Q_t, M)$ as discussed in [6]. He found that the behavior of the bounded models approaches the behavior of the unbounded TPM provided that the capacity constraint of each station is sufficiently large relative to the workload.

Fine and Graves applied the TPM to a real-world job shop that manufactured thermal conduction modules for IBM mainframes [8]. Here, the model was extended to allow consideration of features such as release policies. By using regression methods, the parameters were fitted to the observed data and they found some empirical evidence for the use of linear control rules in practice. The model was then used to study the impact of various planning policies and the effect of changes in product mix.

Hollywood made several extensions to the TPM [9]. He defined the class of MR models of which the features are stated in Section 1. Previously, work on this class of model had been limited to the TPM and direct extensions to the TPM. Consequently, all research assumed that the production is a fixed fraction of work-in-process, and that work arrivals are stationary fluid arrivals. He showed that the TPM is one model in a much larger class of models that may be analyzed through similar techniques.

He expanded the TPM to include models with general linear control rules which are also known as affine control rules. These rules allow production to be a weighted sum of inventory levels at multiple stations, plus a random noise term. He then applied these models on a network that uses highly sophisticated affine control rules, i.e. the proportional restoration rules of Denardo and Tang [10].

Next, he demonstrated how to calculate approximations for the steady-state moments of MR models with general non-linear control rules. These models allow for the modeling of a wide range of realistic machine and human behavior, including machine congestion and effects of overtime work. He suggested a non-linear control rule which is based on Karmarkar's "clearing function" [11]. However, the results are approximations. Errors can be large if the network is heavily loaded and work arrivals are highly variable.

Hollywood also showed how to set up and solve optimization problems related to MR models, including maximum performance (minimum queuing times or queue lengths, and maximum throughput) and minimum cost problem. In doing so, he showed how to model capacity requirements for models with linear or general control rules.

He also used the underlying recursion equations to find the transient behavior of MR models that are subjected to changes in the network. For simple network changes, he found the analytical expression for the transients, extending the work of Parrish [3]. Using this transient analysis, he also found that the optimal control rules are linear functions of the work-in-process, thereby justifying the use of linear control rules.

Graves and Hollywood developed a constant-inventory TPM in which the release of work into the shop is regulated to maintain a constant inventory level [12]. They were able to determine the first two moments for the production random vector for such a release policy and also characterized the conditions for which the production levels converge to a steady state. They then illustrated the use of the model with an application and showed the benefits of such a release policy with a computational experiment.

B. Other Related Models

We now compare the MR models with other related models. Our focus here is to compare the modeling framework and the approaches, rather than the intent of the models.

Queuing models are widely used to model processing networks. Jackson developed the basic queuing model for open queuing networks, now known as the Jackson networks [13][14]. Similar to the MR models, the detailed sequencing of jobs is not considered and work flows can be characterized in a complex job shop. Gordon and Newell extended Jackson's work to a closed queuing system in which the number of jobs remains constant [15]. There is now a large literature on queuing networks that extends and generalizes Jackson's work, with much of it validating against simulation studies. Buzacott and Yao, and Suri and Sanders provided an extensive survey on queuing models [16][17].

Queuing models usually assume the arrival and service times to be independent and identically distributed. In order to compute the exact solutions, the arrivals are assumed to follow a Poisson process with exponentially distributed service times. Whitt provided approximations for general networks with GI/G/m queues using the first two moments of inter-arrival and service times [18][19]. The resulting queuing model is known as the Queuing Network Analyzer. Bitran and Tirupati extended this model by improving the approximation accuracy of the multiple customer class version of the model [20]. They achieved this by developing an improved approach to estimate the interaction between customer classes, albeit it requires more complex computation.

Much recently, there is some literature that focuses on the approximations of heavy-traffic queuing models. Work in this area includes Harrison and Williams, Harrison and Wein [21]-[23]. These are Brownian motion approximations based on the fact that departures from heavily-loaded stations are approximately exponentially distributed. However, the major drawback of these models is that the accuracy of these models largely depends on the traffic-intensities of the queues, and thus they are more applicable to heavily-loaded queues.

MR models allow the modeling of splits and merges in workflows, i.e. a single job can split into multiple jobs, or multiple jobs can merge into a single job upon completion and moves to downstream stations. But this is generally not possible in queuing models. Furthermore, MR models allows the computation of first two moments of both production quantities and queue lengths, while queuing models usually gives only the first moments. However, it is difficult to model probabilistic job routing in MR models which is a basic feature of queuing models, although Graves presented an approximation method for such modeling for the TPM [6]. Generally, MR models are more suitable if the production output per time period can be expressed as a function of the total work and the network has well-defined workflows. Queuing models are more appropriate if the network has stations with independent and identically distributed service times and has distinct classes of "customers".

Besides queuing models, it is also worthwhile to compare the MR models with deterministic planning models that are often used for aggregate and capacity planning. Compared to MR models which are stochastic models, the deterministic models assume the modeling parameters to be deterministic. The expected production quantities and queue lengths are usually assumed to be deterministic quantities. The main advantage of deterministic models over MR models is that it is possible to model complicated production rules such as bounded control rules which is not possible with MR models. However, capacity is usually a hard constraint in deterministic models. A station *i* processes up to its fixed capacity in each discrete time period, i.e. production of station *i* is given by the bounded control rule $P_{it} = min(Q_{it}, M_i)$. As a result, such models do not account for the lead time or work-in-process consequences of capacity loading, as lead time of production (and work-in-process) is constant regardless of the amount of capacity loading. Reviews of these models can be found in Hax, Baker, and Bitran and Tirupati [24]-[26].

One deterministic model that is closely related to the MR models is the Input/Output control by Wight (1970). The Input/Output control is a way of analyzing the consequences of order release decisions of material planning. Similar to the MR models, it is a discrete-time processing network model at a shop level. In this model, work arriving at each station is determined and the amount of work processed is computed by the bounded control rule. Work in excess of capacity is carried over to the next period. The discrete treatment of work order and job step of this model causes it to resemble a discrete-time simulation. Thus the resulting computations required are as detailed and complex as the real simulation, and hence it is difficult to embed it in optimization procedures.

Karmarkar's deterministic model takes into account of the capacity-loading effect on the lead time and work-in-process by a "clearing function" [11]. Similar to the control rules in MR models, the clearing function describes the amount of output "cleared" from the manufacturing facility as a function of its work-in-process. The function is

$$P = \frac{MQ}{\beta + Q} \qquad (4)$$

where P is the production rate, Q is the work-in-process level, M is the capacity level, and β is a parameter that determines the clearing rate. He adapted (4) for a discrete period dynamic planning model that directly models work-in-process and finished inventories. Since it is a deterministic model, it can be incorporated readily in mathematical programming techniques.

Unlike the clearing function, the linear control rule in the TPM does not capture the effect of capacity loading on work-in-process and production lead times. Hence Hollywood suggested a non-linear control rule based on the clearing function and also illustrated an approximation method to compute the steady-state moments of MR models with general non-linear control rules [9].

In this section, we have presented the literature on MR models as well as other related models. However, there still exist some significant research gaps, which will be discussed in the next section.

IV. Research opportunities

There is substantial literature on MR models as presented in the previous section. However, additional work is needed to make MR models more realistic and applicable in real-life manufacturing facilities. There are more unresolved issues than can be listed; we will look at some that seem interesting and valuable.

Many production facilities show congestion effect due to capacity loading. Hollywood has derived the non-linear control rule that is based on Karmarkar's "clearing function" which is a concave function designed to model saturation and congestion behavior [9][11]. But Hollywood's non-linear control rule does not perform well in high traffic-intensity conditions and when the arrival is highly variable. This greatly limits the use of this model since stations of high traffic-intensity (bottleneck stations) are usually the ones that are more crucial, and thus require a more thorough performance evaluation. Hollywood performed a discrete-time simulation study to validate his approximations. He found that errors in approximating the second moments of both production and queue lengths of a single-stage station are larger than 50% when subjected to a traffic-intensity of 0.9 and work arrivals with a coefficient of variability equal to 1.0. Therefore, more effort is required to improve the non-linear control rules or the approximation method.

One feature of the MR models is that work transfers between stations occur only at the beginning of each discrete time period. In many manufacturing facilities, jobs move to the next processing station immediately after completion at each station. Thus there is a continuous shuttle of jobs between stations of the network. Graves suggested that the MR models still applies to such an environment if the periods are carefully sized [1]. The periods must be long enough such that a significant amount of work is done in each period, but short enough that a job is unlikely to travel through more than one station in a particular period.

However, such period-sizing might be restrictive in general applications. In many production systems, it takes only a short time (e.g. less than half an hour) for a job to travel through more than one station, and thus the discrete time period has to be short. In most applications, a longer time period might be more valuable. This is because the MR model outputs, such as the production variance, are more meaningful if they are defined over a longer time. It is usually more useful for a production manager to know the production variance of a shift, rather than over a one-hour period. For example, a manager planning the capacity might use the expression $Capacity = E[P_t] + k\sigma_P$, where $E[P_t]$ is the expected production quantities per period, σ_P is the standard deviation of production quantities, and k is the "safety" factor that converts the standard deviation of the throughput into performance guarantee. k equals to the z-value of standard normal distribution if P_t is normally distributed. Here, it is hard to set the value of k if σ_p is of a one-hour period even if P_t is normally distributed, as it is difficult to determine the amount of performance guarantee needed for each one-hour period. A performance guarantee for a shift would be more meaningful. It is generally complex to determine the production variance of a longer time period given the production variance of shorter time periods. This is because the production quantities of successive periods are generally correlated.

Hollywood's approximation for non-linear control rules is more accurate if the arrival process is less variable [9]. In many other forms of network approximations in which there is a stochastic arrival process, such as Whitt's approximation of GI/G/m queues [18][19], a high level of arrival variability generally worsens the accuracy of the approximation. By having a longer discrete time period, the variability of the amount of work that arrives in each period will be lower due to variability pooling over a longer period of time. Therefore, in the case where an exact solution of the MR model cannot be obtained and thus requires an approximation, a longer discrete time period will probably improve the accuracy of the approximation.

We are in the process of building a model that attempts to fill the above mentioned research gaps. This model accounts for machine congestion and has significantly more accurate second moment approximations than Hollywood's non-linear control rule model. Furthermore, it models continuous flow of jobs and can have discrete time periods longer than the time for a job to travel through more than one station. Simulation studies were carried out to validate the accuracy of the model. Preliminary results of the studies are encouraging.

As mentioned in Section I, one of the main capabilities of the MR model is its ability to determine or approximate the variances of production volume. Machine failure is the main source of unplanned variability in the production floor. Hence, the ability to include machine failures in the MR model will greatly improve its applicability. Graves modeled an unreliable single station in which the station fails with a fixed probability p and the duration of the failure is exactly one period [6]. This is of limited realism compared to actual machine failures and additional effort is therefore required in this aspect of modeling.

REFERENCES

[1] S.C. Graves, “A Tactical Planning Model for a job shop”, *Operations Research*, 34, 1986, pp 522-533.

[2] A.B. Cruickshanks, R. D. Drescher and S. C. Graves, “A study of production smoothing in a job shop environment”, *Management Science*, 30, 1984, 36-42.

[3] S.H. Parrish, “Extensions to a model for tactical planning in a job shop Environment”, S.M. Thesis, Operations Research Center, Massachusetts Institute of Technology, June 1987.

[4] T.Y. Leong, “A Tactical Planning Model for a mixed push and pull system”, Ph.D. program second year paper, Sloan School of Management, Massachusetts Institute of Technology, July 1987.

[5] S.C. Graves, “Determining the spares and staffing levels for a repair depot”, *J. Mfg. Oper. Mgt,* 1, 1988, pp 227-241.

[6] S.C. Graves, “Extensions to a Tactical Planning Model for a job shop”, *Proceedings of the 27th IEEE Conference on Decision and Control*, Austin, Texas, December 1988.

[7] S. Mihara, “A Tactical Planning Model for a job shop with unreliable work stations and capacity constraints”, S.M. Thesis, Operations Research Center, MIT, Cambridge MA, January 1988.

[8] C. H. Fine and S. C. Graves, “A Tactical Planning Model for manufacturing subcomponents of mainframe computers”, *J. Mfg. Oper. Mgt*., 2, 1989, pp 4-34.

[9] J. S. Hollywood, “Performance evaluation and optimization models for processing networks with queue-dependent production quantities”, Ph.D. Thesis, Operations Research Center, MIT, Cambridge MA, June 2000.

[10] E.V. Denardo and Christopher. S. Tang, “Control of a stochastic production system with estimated parameters”, *Management Science*, 43, 1997, pp 1296-1307.

[11] U. S. Karmarkar, “Capacity loading and release planning with work-in-progress (WIP) and leadtimes”, *J. Mfg. Oper. Mgt,* 2, 1989, pp 105-123.

[12] S. C. Graves and J. S. Hollywood, “A constant-inventory Tactical Planning Model for a job shop”, Working paper, January 2001, 30 pp.

[13] J.R. Jackson, “Networks of waiting lines”, *Operations Research*, 5, 1957, pp 518-521.

[14] J.R Jackson, “Jobshop-like queuing systems”, *Management Science*, 10, 1967, pp 131-142.

[15] K. D. Gordon, and G. F. Newell. 1967, “Closed queuing systems with exponential servers”, *Operations Research*, 15, 1967, pp 254-265.

[16] J. A. Buzacott and D. D. Yao, “Flexible manufacturing systems: A review of analytical models”, *Management Science*, 31, 1985, pp 890-905.

[17] R. Suri and J. L. Sanders, “Performance evaluation of production networks”, *Logistics of Production and Inventory*, S. Graves, A. H. G. Rinnooy Kan and P. Zipkin (eds.), *Handbook in Operations Research and Management Science, Vol. 4*, 1993, North Holland.

[18] W. Whitt, “The Queuing Network Analyzer”, *Bell Syst. Tech*., 62, 1983, pp 2779-2815.

[19] W. Whitt, “Performance of the Queuing Network Analyzer”, *Bell Syst. Tech*., 62, 1983, pp 2817-2843.

[20] G. R. Bitran and D. Tirupati, “Multiproduct queuing networks with deterministic routings: Decomposition approach and notion of inference”, *Management Science*, 34, 1988, pp 75-100.

[21] J. M. Harrison and R. J. Williams, “Brownian motion models of open queuing networks with homogeneous customer populations”, *Stochastics*, 22, 1987, pp 77-115.

[22] J. M. Harrison, “Brownian models of open queuing networks with heterogeneous customer populations”, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, W. Fleming and P. L. Lions (eds), IMA Volume 10, Springer-Verlag, New York, 1988, pp 147-186.

[23] L. M. Wein, “Dynamic scheduling of a multiclass make-to-stock queue”, *Operations Research*, 40, 1992, pp 724-735.

[24] A. C. Hax, “Aggregate production planning”, *Handbook of Operations Research, Vol. 2,* J. Moder and S. E. Elmaghraby (eds), Von Nostrand, Reinhold.

[25] K. Baker, “Requirements Planning”, *Logistics of Production and Inventory*, S. Graves, A. H. G. Rinnooy Kan and P. Zipkin (eds.), *Handbook in Operations Research and Management Science, Vol. 4*, 1993, North Holland.

[26] G. R. Bitran and D. Tirupati, “Hierarchical Planning”, *Logistics of Production and Inventory*, S. Graves, A. H. G. Rinnooy Kan and P. Zipkin (eds.), *Handbook in Operations Research and Management Science, Vol. 4*, 1993, North Holland.

[27] O. Wight, “Input/Output Control – A real handle on lead time”, *Production & Inventory Management*, 3rd Qrt.