Modelling short-term manufacturing flexibility by human intervention and its impact on performance

Modelling short-term manufacturing flexibility by human intervention and its impact on performance

Ton G. de Kok

School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Paviljoen E.12

P.O. Box 513

5600MB Eindhoven

The Netherlands

a.g.d.kok@tue.nl

Ton de Kok is professor in Quantitative Analysis of Operational Processes at TUE since 1992. He received his BSc in Mathematics and Physics and his MSc in Mathematics from Leiden University, and his PhD from Free University in Amsterdam. He spent 8 years in industry with Philips Electronics. His research interests comprise multi-echelon inventory theory, queueing theory, production planning, vehicle routing, and hierarchical planning. He is ISIR Fellow since 2014.

Modelling short-term manufacturing flexibility by human intervention and its impact on performance

Short-term manufacturing flexibility is the capability to respond to unforeseen short-term and immediate events during operation. Explicitly modelling short-term manufacturing flexibility is extremely hard as it implies that it is possible to prescribe responses to all possible events in all possible states of the system under consideration. We propose an implicit modelling approach where we only model the frequency of responses to unforeseen events and the impact of the response on the evolution of the process under consideration. This allows us to dramatically reduce model complexity while preserving empirical validity of the model. We introduce the concepts of intervention-independent performance (IIP) and intervention-dependent performance (IDP). We present a methodology to gather IIP and IDP indicator information. We propose another process to use this information together with historical transactional data and forecasts to make tactical trade-offs that result in control policy parameters. We illustrate our modelling approach by a generic inventory control model with replanning of replenishment orders. We apply the model in a case study characterized by highly utilized production lines, high set-up times and dynamic demand. We provide empirical evidence of the validity of the proposed methodologies.

Keywords: flexibility, performance measurement, rescheduling, inventory management, human intervention

1 Introduction

Flexibility in manufacturing has been subject of extensive research since the mid 1980s, when it was identified as a key means for business success (cf. Bertrand (2003), Beach et al. (2000), and Jain et al. (2013)). Most authors conclude that manufacturing flexibility has not been defined conclusively. We follow Olhager (1993) in defining short-term manufacturing flexibility as the ability to adapt to changing conditions using the existing set and amount of resources. The ability to adapt depends on the capabilities of the human planners and schedulers, the information systems they use, and the pro-active investment in slack resources (i.e. safety stocks, excess processing capacity, safety times). In Barad and Sipper (1988) various flexibility types are defined that affect short-term manufacturing flexibility: machine set-up flexibility, process flexibility, transfer flexibility, system set-up flexibility, and routing flexibility. They use Petri nets to model the various flexibility types quantitatively. Barad (2013) distinguishes between a bottom-up perspective where elements of flexibility are modelled, and a top-down perspective, where flexibility is seen as strategic instrument. Our paper fits into the bottom-up perspective, though we do not explicitly define the type of short-term manufacturing flexibility considered.

In this paper we aim to *quantitatively* model the impact of short-term manufacturing flexibility on manufacturing system performance, measured in both customer service and capital invested in on-hand inventory, without being specific on the way this flexibility has been created. This quantitative approach compels to *measure* short-term flexibility and to relate the flexibility *level* to the performance of the manufacturing system. We empirically validate our quantitative modelling approach in a real-life

situation, where we use an analytical model in conjunction with discrete event simulation.

The International Journal of Production Research has been a prominent platform for advancing the knowledge on operational (or manufacturing) flexibility. We found 126 papers published in IJPR on the subject. Seebacher and Winkler (2013) found that IJPR published most papers on flexibility in comparison with other journals. In their extensive discussion of literature until the late 1990s, De Toni and Tonchia (1998) conclude that "the measure of flexibility is still an under-developed subject, both for its multi-dimensionality and the lack of indicators for its direct measurement". In this paper we propose the *replanning frequency* as a directly measurable flexibility indicator. Zhang and Tseng (2009) introduce customer flexibility as a means to reduce costs in a high mix low volume environment. Customer flexibility is key in such circumstances, as holding inventory in such environments is not an economically viable option. We found that customer flexibility is also of importance in high-volume capacity constrained environments. In our paper we model customer flexibility explicitly assuming a positive customer lead time. In their extensive survey of more recent publications on flexibility, Jain et al. (2013) discuss a wide range of flexibility measures. When considering these measures in more detail, they represent *a priori* measures of flexibility, i.e. (relatively) high scores on these measures indicate the capability to be flexible. However, these measures do not capture the actual *exploitation* of flexibility. The replanning frequency is such an *a posteori* measure of flexibility.

The added value of human planners is their capability to identify contingencies that enable to mitigate the impact of unforeseen events. To some extent human planners find

unique solutions to unique problems caused by a unique combination of events. This observation implies that it is impossible to model the behaviour of the planner at the detailed level of causal chains of events. In contrast, most quantitative models assume some standard response to events. In the context of scheduling typical standard responses to customer order arrivals are FCFS or Operation Due Date scheduling. In the context of inventory management, typical responses are reorder when the (echelon) inventory position is below the reorder point or reorder at the start of next week. Such standard procedures fail to model the contingencies, or inherent flexibility options, that are typical for real-life multi-item multi-resource environments. By definition, actual actions taken depend on the specific state of the system to be managed and in most cases create dependence between on the one hand the demand process and on the other hand the production process. Analysis of stochastic models with high-dimensional state spaces and dependence between different stochastic processes is infamous for its complexity. Conversely, in the presence of short-term manufacturing flexibility, resorting to application of standard inventory and queueing models may yield far too pessimistic estimates of actual performance.

Here we propose an approach to avoid the curses of dimensionality, while taking short-term manufacturing flexibility into account. Instead of modelling the actual actions and the resulting process evolution over time, we only model the *effect* of the interventions of the human planners and the *intervention frequency*. Though we illustrate our approach by a specific inventory management problem, we expect that the approach is generic and a possible route to apply basic OM models in a real-life context. In order to model the *effect* of interventions of human planners, we must *measure* this effect using transactions systems, such as Enterprise Resource Planning (ERP) systems and

Manufacturing Execution Systems (MES). This measurement must be such that it measures both the performance of the system without (or before) the human intervention, and the performance with (or after) the human intervention. This leads to the concepts of *intervention-independent performance* (IIP) and *intervention-dependent performance* (IDP). The approach illustrated in this paper can only be applied if both performances can be measured in quantitative terms and the system without human interventions can be quantitatively modelled.

Short-term manufacturing flexibility under stochastic demand and supply is primarily discussed in the context of spare part systems. Verrijdt (1997) proposed a short-term manufacturing flexibility framework for spare parts systems. Repairable items pipeline flexibility is modelled in Dada (1992) and Verrijdt et al. (1998). Pooling of spare parts inventories by allowing for transhipments is discussed in Axsäter (1990, 2003). We refer to Van Houtum and Kranenburg (2015) for an extensive overview of short-term manufacturing flexibility in service supply chains.

In the context of product supply chains Simpson (1958), and Inderfurth and Minner (1998) study multi-echelon inventory systems with short-term manufacturing flexibility under uncertainty. They assume that any demand causing a stockout at some stockpoint is satisfied by some implicit flexibility, e.g. by expediting or outsourcing. The main difference between this modelling approach and ours is that we explicitly measure the frequency of human interventions in response to (potential) stockouts and incorporate this frequency into the customer service level determined by the mathematical model. We assume that the frequency of interventions is an explicit measure of flexibility. Related to our work is the model discussed in Minner et al. (2003). They study a two-

echelon supply chain, where it is assumed that in case of stockouts at the warehouse, one or more outstanding warehouse orders can be expedited with some probability to be received immediately. The assumption of a probabilistic mechanism underlying the expedition of an item work order or not is similar to the assumption underlying our modelling: from the point of view of a single sku the selection of its work order to be expedited is contingent on the status of other items' work orders and inventory availability, whereby no explicit mechanism other than a probability of success can be inferred.

The contribution of our paper is as follows. Firstly we provide an extensive description of the practice of planning and scheduling under uncertainty based on intense collaboration with industry over a period of three decades. This leads to the introduction of the concept of intervention-independent and intervention-dependent performance measures. We apply these concepts to a generic single-item single-echelon inventory model and derive expressions for customer service levels. Finally we provide empirical evidence of the applicability of our modelling approach in the context of the real-life case, which is characterized by multiple end-items produced on a number of highly utilized and inflexible production lines.

The paper is organised as follows. In section 2 we discuss the concept of intervention-dependent and intervention-independent performance indicators and argue that only the latter ones can be used for application and calibration of mathematical models. In section 3 we describe the mathematical model considered. In section 4 we present a heuristic analysis of the model. In section 5 we apply the model to a real-life case, thus

providing empirical evidence of the validity of the approach. In section 6 we present conclusions and ideas for further research.

2 Performance indicators and human interventions

In this section we provide a description of responses to unforeseen events enabled by information about the status of item work orders and short term item availability. We distinguish between the Make-To-Stock situation where customers expect immediate availability and situations where customers expect some nominal positive customer order lead time. We argue that the appropriate human interventions can substantially improve customer service, but that performance of supply chains with human interventions is mathematically intractable. As we need mathematical models to set control parameters, such as safety stocks, safety times and nominal lead times, we need to identify supply chain performance *without human intervention* from transactional data. On the one hand this enables formal tactical trade-offs to set control parameters, and on the other hand enables calibration of performance targets *before human intervention* in relation to performance targets *after human intervention*.

In supply chains we distinguish between two operational management processes, namely Goods Flow Control (GFC) and Production Unit Control (PUC) (cf. Bertrand et al. (1990). GFC accepts customer orders and coordinates the order releases to production units (PUs), which process the orders in such a way that the due dates agreed with goodsflow control are met with a high probability. By doing so, GFC may consider the production units as a black box with a constant delay. GFC operationally coordinates the supply chain of a company from purchasing to sales.

However, both GFC and PUC are confronted with unforeseen events. GFC must cope with uncertainty in customer demand with respect to timing, quantity and specification, and uncertainty in lead times of external suppliers. PUC must cope with machine break-downs, setup times, and work order releases that differ from expectations earlier. The production unit exploits its flexibility created by slack capacity and time, and dynamically controls work order priorities to ensure alignment with their due dates, e.g. by using an Operation Due Date (ODD) dispatching rule at each work station. We note here that flow times in a PU under ODD rules are not mathematically tractable. However, if PUC ensures a high due date reliability, we do not need to bother about that.

At GFC level unforeseen changes are absorbed by time and material buffers. At the Customer Order Decoupling Point (CODP), (cf. Hoekstra and Romme (1992) and De Kok and Fransoo (2003)), i.e. the echelon that decouples the forecast-driven part of the supply chain from the order-driven part, customer orders may require items that are not available in stock. Here we must distinguish between Make-To-Stock situations and other situations, where the CODP is not at end item level, as is the case for Assemble-To-Order (ATO), Configure-To-Order (CTO) and Make-To-Order situations (MTO).

(i) Make-To-Stock, no advance demand information

Under MTS without advance demand information the customer expects immediate availability. As this is not the case, the customer may decide to cancel its order and its demand is lost. Or the customer accepts a later delivery moment as derived from Available-To-Promise (ATP) information. It may be that the PU producing the item is contacted to expedite the item work order to ensure timely delivery according to the agreed delivery moment.

Note that in case of lost demand we need not change order priorities, but we must register the lost sale. If the customer is willing to wait, we see two events occurring at the same time: an item work order is expedited and an order delivery is postponed. This creates correlation between the demand process and the supply process that is hard to analyse mathematically. The mismatch between demand and (planned) supply can be identified at the moment the customer ordered the item. Such a mismatch is similar to the mismatch between demand and supply in classical inventory models. The potential stockout occurred before human intervention that created correlation between demand and supply. We assume that the potential stockout results from two independent processes, as we assume that the original item order lead time was according to a preset norm.

(ii) Make-To-Stock, advance demand information

Under MTS with advanced demand information end-item work orders are released before customer orders are known, yet there is an agreed Customer Lead Time (CLT) between the order moment and delivery moment. This situation is quite common in capacity-constrained high-setup-time environments upstream in most supply chains. Typically in such environments a production wheel is used to minimize setup times. Due to large batch sizes substantial cycle stocks are created at the CODP from which routinely demand can be satisfied. However, occasionally a customer order cannot be delivered from stock due to depletion of end-item stock and no work order planned within the customer lead time. In that case the production wheel is reconsidered in order to rebalance inventory with future demand forecasts. Here GFC provides information about desired short-term or mid-term planned on-hand availability and PUC derives

from that work-order quantities to be produced. PUC uses its (limited) flexibility, the customer order information and the required planned on-hand availability to create a new schedule that eventually should enable the PU to both satisfy the customer orders and to get back to the production wheel rhythm as soon as possible. Again we observe that a customer order triggers a complex set of events in the PU that yields flow times that are impossible to derive mathematically. Again we can state that before the human intervention we identified a mismatch between demand and supply that is similar to the mismatch in classical inventory models, such as the multi-item stochastic economic lot-sizing problem (SELSP) (cf. Smits et al. (2004)).

(iii) Not Make-To-Stock

Under ATO, CTO and MTO the customer order initiates two events:

a. Consumption of all items at the CODP needed for the customer order to be assembled.
b. Release of a work order to the PU assembling the customer order.

If all items are available, the work order can be released according to its nominal customer lead time, assuming that the customer accepts this lead time. If some items are not available GFC needs to communicate with the internal PU upstream of the CODP that produces these items or with outside suppliers to ensure expedited delivery. In parallel it communicates with the PU downstream of the CODP that assembles the customer order to communicate that this order needs expedited assembly as well, as soon as all items are available. In both PU's work orders must be rescheduled to enable expedited delivery, while some other work orders are delayed. In the upstream PU the latter may be work orders

for items with sufficient short-term availability at the CODP. In the downstream PU work orders which are ahead of schedule may be delayed.

Again, this interaction between GFC and PU's is mathematically intractable: orders are expedited and delayed given their due dates and given availability of items in stockpoints. But the initial imbalance between demand and supply, and consequential out of stocks, can be captured by multi-item multi-echelon inventory systems under stochastic demand and nominal lead times.

In all of the above situations we are confronted with human interventions that restore the balance between demand and supply by expediting work orders and, if needed, delaying customer orders. In order to do so, detailed real-time information on the status of *all work orders in a PU* is used, whereby the state space of the system is extremely large. Human planners and schedulers only implicitly are aware of this, as they developed routines to cope with imbalances between demand and supply: their mental models. We postulate that it is impossible to formally model both the state space, the action space and the transition probabilities that allow optimization using stochastic dynamic programming. Though planning and scheduling algorithms can support the planner and scheduler, these algorithms usually assume deterministic demand and processing times, and do not capture the tactical trade-offs to be made that create the required flexibility in the first place: trade-offs between customer service, capital invested in equipment and inventory, and costs of labor and equipment usage, under uncertainty in demand, production and supply.

The impact of human interventions on system performance can be huge compared with the performance derived from mathematical models based on transactional data on

supply and demand. We came across situations where customer service levels computed from classical inventory models were below 90%, while actual service was above 95%. Work order lead times may show high volatility, which translates in high safety times and stocks when using mathematical models, while this volatility may be deliberately created in response to unforeseen events, and does not affect item availability downstream of the PU under consideration. We propose to rely on nominal lead times used in planning, as the resulting due dates are the targets to be met. Thus blackboxing the PU's simplifies control at GFC level, while leaving the exploitation of flexibility to the PUC level. In De Kok (2015) evidence is provided of the empirical validity of multi-item multi-echelon inventory systems with nominal lead times to explain the actual customer service from average item inventories, stationary i.i.d. demand and average lot sizes. Essential to this empirical validity is the availability of customer service performance that is not impacted by human interventions. Above we already indicated how such information could be gathered. Let us describe this in more detail with the support of figure 1.

Firstly, we assume that norms for nominal customer order lead times, work order lead times, lot sizes, safety times and safety stocks have been determined. Secondly, we assume that GFC and PUC policies are in place. The GFC policies allow for routine work order releases to the PU's. The PU policies allow for routine work order execution in line with due dates set. At GFC level routine customer order acceptance policies are in place as well, such as FCFS or priorities according to some customer classification scheme.

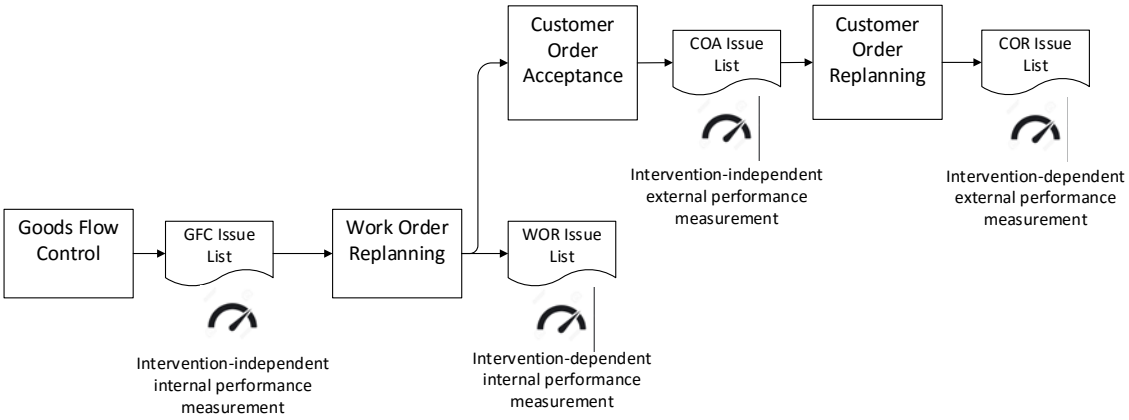

Figure 1. Measuring intervention-independent and intervention-dependent performance

The GFC process is periodic, typically weekly. Due to forecast updates and information about actual work order completions that may differ from the planned work order completion according to last week's plan, issues may be identified, such as material and capacity shortages, both short-term and mid-term. All these issues should be reported in the form of a GFC issues list. From this list we can determine the internal IIP indicators related to material availability, such as item fill rates and item non-stockout probability, and IIP indicators related to capacity availability, such as probability of capacity overload per resource type. The resulting immediate and planned order releases are communicated with PUC and work orders are replanned in order to restore the balance between demand and supply. This results into a Work Order Replanning (WOR) issue list from which the internal IDP indicators can be determined, such as due date reliability, resource utilization and overtime. We note here that we should not only measure the planned performance after replanning, but also the frequency of replanning, as this provides us with valuable information about the effort spent on replanning. Similarly it is valuable to record additional costs incurred due to replanning, e.g. inefficiencies in processing, outsourcing, expedited shipments. Note that we focus on the GFC level, as here the balance between demand and supply is managed. At PUC

level, only due date adherence is relevant. Still, the PUC enables replanning by providing relevant information about the current capability to reschedule particular work orders, and executing the work orders according to the new due dates agreed.

After the work order release decisions have been taken, the resulting ATP is communicated to the order desk. Order acceptance is a real-time process and manages external performance to the customers. Each customer order accepted is documented regarding its arrival time, nominal due date and agreed due date after negotiation with the customer and possible rescheduling of work orders. The IIP indicators are based on the nominal due dates and the ATP before negotiation. Typical examples of external IIP indicators are fill rate and non-stockout probability. The IDP performance is based on agreed delivery dates and actual delivery dates. A typical example of an IDP indicator is fraction of confirmed orders delivered on time. The IDP performance should be close to 100% for the accepted customer orders, but can overall be substantially below 100% due to lost-sales. The IIP performance depends on the particular situation. We have come across situations that an IIP performance of 80% in terms of on-time delivery is acceptable. Striving for higher IIP targets would be economically infeasible, and unnecessary when IDP performance is up to standard.

In figure 2 we show how IIP and IDP indicators can be used to derive control policies for future use. We focus on the GFC problem, but a similar approach can be applied to develop models for the PU problem, e.g. queueing network models. The GFC problem of coordination of work orders across PU's and customer order acceptance can be described by multi-item multi-echelon (MIME) inventory models. Transactional data concerning work order lead times, actual demand, actual lot sizes and actual inventory

levels can be used as input for these models. Current control policies can be used as well, but we should be cautious as in most cases the formally described control policies are often not adhered to. Instead we can infer the control policies from the average inventory levels and average lot sizes computed from the transactional data.

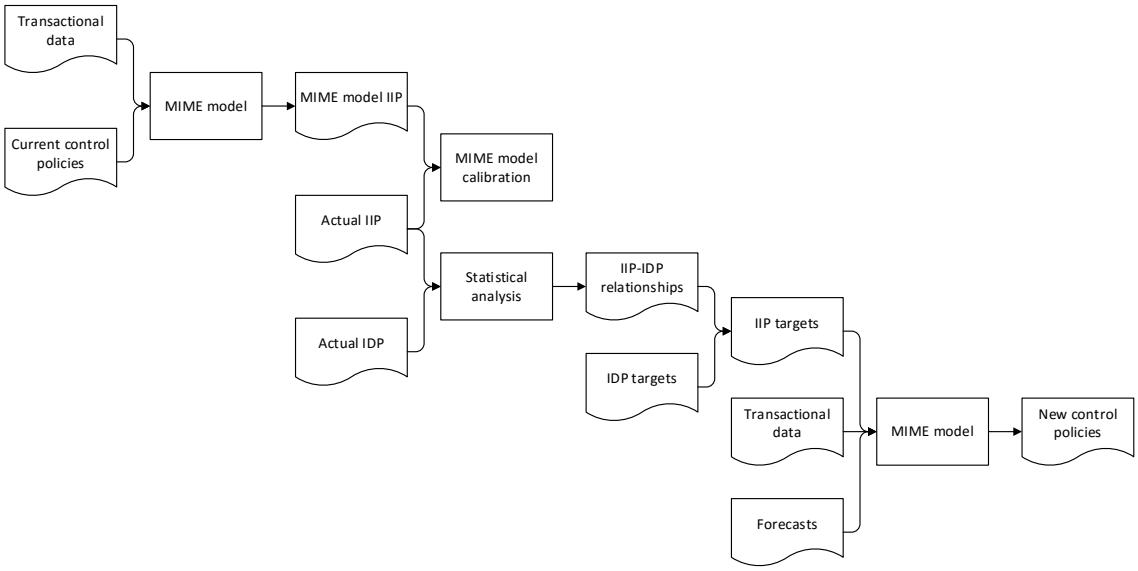

Figure 2. Empirical validation and calibration of mathematical model to derive control policies for future use

By comparing the model IIP indicators with the actual IIP indicators we can check empirical validity of the MIME model, and if needed we can calibrate the MIME model to improve the correspondence between model and actual IIP indicators. Clearly such calibration is an art that should be done with care and understanding. Using statistical modelling we can infer the relationship between actual IIP and actual IDP indicators. Note that IDP indicators should include frequency of replanning as a measure of flexibility. We expect that a higher planning frequency leads to higher intervention-dependent customer service for the same intervention-independent customer service. Note that in global supply chains many end-items across many stockpoints provide

actual IDP indicators that can be compared against actual IIP indicators. Next we can set future IDP targets from which the IIP targets can be derived, using the statistical model. Finally we can use the MIME model to derive new control policies and their parameters for future use. As inputs we can use forecasts on demand, but also nominal lead times determined from new targets for the PU, possibly derived from a PUC model. We probably need historical demand data to determine reasonable values for forecast errors and coefficients of variation.

This concludes our discussion of planning and control processes in real-life situations, the role of the human planner and scheduler, and the importance of the distinction between IIP and IDP indicators for the use of mathematical models in practice. In the next sections we present a mathematical model and its application to a real-life case to illustrate the viability of the proposed approach.

3 Model description

We illustrate our approach by considering a single PU multi-item MTS situation with a positive customer lead time (cf. De Kok (2011)). For each item an (*R, s, Q*)-policy is used as routine control rule. I.e. at the beginning of each review period of R time units the inventory position is monitored. If the inventory position is below s, we release a work order consisting of a multiple of *Q* items to the PU, such that the item inventory position is raised to a value between s and *s*+*Q*. If the inventory position is above *s* no order is initiated. The normal production lead time equals *L*, where *L* is a multiple of *R*. Without loss of generality we assume *R*=1. This implies that each production order initiated at the beginning of review period *t* is assumed to be produced in period *t*+*L*. Here period *t* is the interval (*t*-1, *t*]. We assume that customer orders arrive at the beginning of period *t*. These orders must be delivered in period $t+ L_c$ with $0 \leq L_c \leq L$, i.e. L_c is the customer lead time. We assume that customer demand in subsequent periods is i.i.d.. Let *D* denote the demand for the item in an arbitrary period. Furthermore we define *D*(*t*,*t*+*n*] as the cumulative item demand during the time interval (*t*,*t*+*n*]. At the beginning of period *t* the following quantities are determined.

$I_t(t+n)$:= forecast of on-hand stock at the start of period *t*+*n*, $n \geq 0$.

$p_t(t+n)$:= forecast of amount produced during period *t*+*n*, $n \geq 0$.

$\mu_t(t+n)$:= forecast of demand during period *t*+*n*, $n \geq 0$.

Note that the index *t* in the above definitions emphasizes the fact that at the start of period *t* forecasts are made of future on-hand stocks, production quantities, and demand.

It follows from the definition of $\mu_t(t+n)$, D, $D(t,t+n)$ and L_c that

$$\mu_t(t+n) = \begin{cases} D(t+n-1, t+n]; & 0 \le n \le L_c \\ E[D] & n > L_c \end{cases}$$

Thus, the forecast $\mu_t(t+n)$ equals the known demand within the customer lead time L_c, and equals the average demand $E[D]$ beyond the customer lead time, as we assume demand is i.i.d.. Furthermore we have

$$I_t(t+n+1) = I_t(t+n) + p_t(t+n) - \mu_t(t+n), \mathrm{n} \ge 0$$

Here $I_t(t)$ is the actual stock at the beginning of period t, i.e. at time t-1. The planner observes $I_t(t+n), \mathrm{n} \ge 0$. If $I_t(t+n) < 0$ for some n, the planner may be able to replan $p_t(t+m), m \ge 0$, such that $I_t(t+n) \ge 0$ after replanning.

Define

π := probability that a replanning action is decided upon.

A successful replanning action eliminates negative $I_t(t+n)$. Given the model parameters R, s, Q, L, L_c and π we want to determine the service levels P_1 and P_2 achieved,

P_1 := the probability of a stockout at the end of a replenishment cycle

P_2 := long-run fraction of demand satisfied in time.

Here a replenishment cycle is defined as the time between arrival of two consecutive replenishments. Once we have an expression for P_1 and P_2 as a function of the model parameters we can conversely determine e.g. s^* as a function of the other model parameters, such that P_1 (P_2) is equal to a target value P_1* (P_2*). We claim that through the introduction of π a target value of 1 can be realised with a finite average on hand stock. This contrasts with results for the classical single echelon one-product inventory models like (*s*,*S*), (*s*,*nQ*), (*R*,*s*,*S*) and (*R*,*s*,*nQ*) for e.g. normally distributed or gamma distributed lead time demand (cf. Silver et al. (1998)).

4 Analysis of the model

In this section we derive expressions for P_1 and P_2 as a function of the model parameters (cf. De Kok (2011)). The analysis consists of two steps. First of all we give an expression for P_1 and P_2 in case no replanning is possible. This is equivalent to the classical inventory model assumption. Secondly we use a number of assumptions through which we can use the results for the model without replanning to derive an expression for P_1 and P_2 in the model with replanning. To make a distinction between characteristics for the model with and without replanning we use upper indices (*r*) and (*nr*), respectively. Thus,

Let us first assume that replanning is not possible. In that case we have a classical (*R*,*s*,*Q*)-policy with the only difference that we have to account for the fact that at the beginning of period *t* the future demand $D(t,\ t+L_c]$ is exactly known. Hence we have only uncertainty in lead time demand during $(t+L_c, t+L]$. In case L_c=0 the appropriate control variable is *Y*(*t*),

$Y(t)$:= inventory position at the beginning of period t.

In case L_c>0 we propose to use W(t) as a control variable,

$W(t)$:= $Y(t) - D(t, t+L_c]$

and to adapt the (R, s, Q)-policy such that an order of Q is initiated when $W(t) < s$. Here we assume that the undershoot of s is small compared to Q, implying that the order quantity equals exactly Q. In that case it can be shown that

$$P_1 = P\{D(L_c, L] + U > s\}$$

$$P_2 = \frac{1 - E[(D(L_c, L] + U - s)^+] - E[(D(L_c, L] + U - (s+Q)^+]}{Q}$$

Here

U := undershoot of reorder level s.

By fitting mixtures of Erlang distributions to the first two moments of $D(0, L - L_c] + U$ (cf. Janssen et al. (1999), we can numerically obtain an excellent approximation for P_1 and P_2.

Now let us assume replanning is possible and π is given. Then we observe the following:

The number of stockouts in the model without replanning equals the number of potential replanning actions in the model with replanning.

Here we assume that Q is large compared to $E[D]$, so that replanning a quantity Q at the beginning of period t eliminates all planned stockouts in period $t+n$, $0 \le n \le L$, *i.e.* $I_t(t+n) \ge 0$.

The number of stockouts per time unit in the model without replanning equals the number of replenishment cycles per time unit multiplied by the probability of a stockout at the end of a replenishment cycle. We assume that a replanning action eliminates a stockout at the end of a replenishment cycle. To give an expression for the number of stockouts in the model with replanning we define the following quantities.

$N^{(r)}$:= the number of stockouts per time unit in the model with replanning

$N^{(nr)}$:= the number of stockouts per time unit in the model without replanning

Then we find the following relations

$$P_1^{(nr)} = P\{D(L_c, L] + U > s\}$$

$$P_1^{(r)} = P_1^{(nr)}(1-\pi)$$

$$N^{(r)} = P_1^{(r)} \frac{E[D]}{Q}$$

$$N^{(nr)} = P_1^{(nr)} \frac{E[D]}{Q}$$

Note that $Q / E[D]$ equals the average replenishment cycle length. The above relations imply that

$$P_1^{(r)} = (1-\pi)P\{D(L_c, L] + U > s\}$$

$$N^{(r)} = \frac{(1-\pi)}{Q} E[D]P\{D(L_c, L] + U > s\}$$

From $N^{(r)}$ we can derive an expression for the key performance indicator f_r, which is defined as

$f_r :=$ the number of replanning actions per time unit.

It is easy to see that

$$f_r = N^{(nr)} - N^{(r)},$$

and thus

$$f_r = \frac{\pi E[D]}{Q} P\{D(L_c, L] + U > s\}$$

In order to find an expression for $P_2^{(r)}$ we observe that the average shortage at the end of a replenishment cycle in the model without replanning equals $(1-P_2^{(nr)})Q$. The average shortage at the end of a replenishment, given that a shortage exists then equals $\frac{(1-P_2^{(nr)})Q}{P_1^{(nr)}}$. We assume that in the model with replanning this conditional expectation of a shortage at the end of a replenishment cycle is the same as in the model without replanning. Then the unconditional expectation of the shortage at the end of a replenishment cycle in the model with replanning equals $\frac{(1-P_2^{(nr)})}{P_1^{(nr)}}Q \cdot P_1^{(r)}$.

Thus we find

$$(1-P_2^{(r)})Q = (1-P_2^{(nr)})Q \cdot \frac{P_1^{(r)}}{P_1^{(nr)}},$$

which yields

$$P_2^{(r)} = 1 - \frac{P_1^{(r)}}{P_1^{(nr)}}(1-P_2^{(nr)}).$$

Using the above derived expressions for $P_1^{(r)}$, $P_1^{(nr)}$ and $P_2^{(nr)}$, we obtain

$$P_2^{(r)} = 1 - \frac{(1-\pi)}{Q}\left(E[(D(L_c, L] + U - s)^+] - E[(D(L_c, L] + U - (s+Q))^+]\right)$$

In practice it is often easier to measure f_r than to determine π. Therefore we write $P_2^{(r)}$ as

$$P_2^{(r)} = 1 - \left(1 - \frac{f_r . Q}{E[D] P\{D(L_c, L] + U > s\}}\right)\left(E[(D(L_c, L] + U - s)^+] - E[(D(L_c, L] + U - (s+Q))^+]\right)$$

Finally we assume that the replanning actions do not impact the average stock on hand. This yields

$$\lim_{t\to\infty} E[I(t,t)] \approx s - E[U] - E[D(L_c, L]] + \tfrac{Q}{2}.$$

This completes the heuristic analysis of the model. In the next section we discuss the application of the model in a real-world situation (cf. De Kok (2011)).

5 Application to the glass bulb case

A glass bulb factory produces 110 different glass bulbs for ordinary lamps. There are 4 production lines, of which one is only producing the most common glass bulb, while the other lines produce a mix of different glass bulbs. Each production line produces approximately 6 glass bulbs per second. Due to the characteristics of glass itself and the characteristics of the production process, changing over from a production run of one product to a production run of another product is time consuming. Change-over times range from several hours to several days. Over a range of years the following manufacturing, planning and control characteristics have been empirically established. The manufacturing throughput time ranges from 2 to 6 weeks, dependent on the item. Customer lead time equals 2 or 3 weeks, dependent on the item. This results in a mixed

make-to-stock/make-to-order situation. Manufacturing lot sizes are more or less fixed and in principle a cyclical schedule (production wheel) is used: glass bulb types are produced in a fixed sequence in order to minimise set-up time. A finished goods stock of about one month production is held to buffer against uncertainty in demand during the manufacturing throughput time. This buffer is mainly used to enable the cyclical "minimum change-over time" schedule. Now and then a customer order is received that cannot be filled from stock nor from future planned orders. In that case a replanning process starts, where the available stock is analysed in detail and the existing production plan is adjusted in order to fill all future customer orders. Implicitly information about future unknown customer orders, derived from the long history of collaboration with a relatively small number of customers, is taken into account in this replanning process. The overall result of planning, control and execution of the plans over the last 10 years has been a 100% customer service level.

We consider the above case typical for factories producing items for other factories. In many cases the customer delivery time is shorter than the manufacturing throughput time. This implies uncertainty in demand during the throughput time and therefore finished goods stock is required.

The problem we are faced with in this particular case is the fact that 100% customer service is realised. To our knowledge application of almost any classical inventory control model to this situation would result in an infinite finished goods stock, unless we assume that demand over a finite period is bounded. E.g. assuming normally distributed lead time demand and a P_1-service measure (cf. Silver et al. (1998)) yields a safety factor of ∞, because the safety factor k should satisfy $\phi(k)=1$. Apparently this is not the case in this real-world situation. When assuming bounded demand, 100%

customer service can be guaranteed, however this is likely to result in prohibitively high physical stocks.

Key to the realisation of 100% service level are two aspects. Firstly, customer order lead time is positive, which gives the possibility to anticipate possible disservice. Secondly, in a multi-product situation one usually has some manufacturing flexibility by which possible stockouts for some items can be prevented by replanning without the expense of causing stockouts of other items. The combination of cyclic schedules and replanning actions in case of future stockouts yields a complex multi-item inventory problem. A related model has been studied by Fransoo et al. (1995). We decided to decompose the problem into single-item models. The multi-item aspect is dealt with through the measurement of f_r, the number of replanning actions per time unit, as a measure for flexibility. We measured f_r by analysing together with the planner the production plans of 9 consecutive weeks. A replanning action was identified by the difference in the production planned for an item for the same week in two consecutive production plans. We carefully identified with the planner the cause of the replanning action. This was very important, since a stockout for one item typically causes changes in future production orders for several items. Hence there is a great danger of double-counting. Through this process we identified 30 replanning actions. Assuming that all 110 products have equal stockout probability we derived that

$$f_r = \frac{30}{(9x110)} / week \approx 0.03 / week$$

Since a 100% fill rate $P_2^{(r)}$ is realised, we conclude that $P_1^{(r)} = 0$. This yields

$$0 = P\{D(L_c, L] + U > s\} - \frac{f_r Q}{E[D]}.$$

Thus the reorder level s can be determined from

$$P\{D(L_c, L] + U > s\} = \frac{f_r Q}{E[D]}.$$

For each item we used historical data and information from the planner to determine Q, $E[D]$, $\sigma(D)$, L_c and L. We implemented our heuristic into a discrete event simulation program. Reason for using discrete event simulation was the dynamics of the model due to the fact that customer demand during $D(t, t + L_c]$ is known at time t, and the availability of short-term demand forecasts and associated forecast errors for periods beyond $t + L_c$. We modified the stationary demand model with fixed reorder level into a stochastic dynamic demand model with dynamic reorder levels. We derived the dynamic reorder levels by computing the first two moments of $D(t + L_c, t + L]$ using the demand forecast information. Finally for the last year we collected data about the average on hand stocks for each production group, i.e. the sum of on hand stocks for all items produced at the same production line. In the table below we compare the results of the simulation with the actual data.

production group	average on hand stock (million glass bulbs)	
	Simulation	actual
1	14.0	33.7
2	25.6	26.7
3	32.4	35.7
4	6.8	8.1
total	78.8	104.2

Table 1.Comparison of the model results for stock on hand with actual data (from De Kok (2011))

For production groups 2, 3 and 4 we find acceptable results. For production group 1 we see an unacceptably large difference between the model result and the actual result. Product group 1 in fact consists of only one product: the most fast moving glass bulb (used in households most often). Due to its importance for the total turnover of the glass bulb plant, management had decided to maintain a so-called strategic stock of about 20 Million glass bulbs in order to guard against extraordinary accidents like a glass oven breaking down. Based on this and the acceptable results for production groups 2, 3 and 4 we concluded that our approach was valid. We emphasize here that the complexity of the actual situation is extremely high: demand orientations, replanning, customer lead times. Even adding each of these elements in isolation to a standard inventory model creates mathematical intractability. Our blackbox approach based on *measurement of the impact of human interventions* apparently resolves this intractability, even when a combination of complicating factors is added to the basic model.

6 Conclusions and further research

In this paper we discussed in detail the nature of short-term manufacturing flexibility in practice and the role of human planners and schedulers to exploit this flexibility. We argued that explicitly modelling short-term manufacturing flexibility yields stochastic dynamic programming problems that suffer from the curse of dimensionality and therefore cannot be solved. As an alternative, we propose a generic mathematical modelling approach to implicitly take short-term manufacturing flexibility into account. We introduced the notions of Intervention-Independent Performance (IIP) and Intervention-Dependent Performance (IDP) indicators. We argued that IIP indicators can be used to validate and calibrate mathematical models. By identifying the relationship between an IIP indicator and its associated IDP indicator, e.g. by statistical analysis or by directly measuring auxiliary process indicators, we can derive target IIP values from target IDP values. The target IIP values can be used in the mathematical model of the system under consideration to set control policies and their parameters.

We illustrated the approach by extending the basic (R,s,nQ)-model with the concept of replanning frequency, which is an example of an auxiliary process indicator. The replanning frequency is the measure for short-term manufacturing flexibility. In the glass bulb case study we used the (R,s,nQ)-model with replanning frequency to explain the average inventories needed to meet 100% customer service after human intervention. Thus we provided evidence that the proposed implicit modelling of short-term manufacturing flexibility is empirically valid.

The specific case studied should be seen as a typical example of how human interventions can be taken into account with tractable quantitative models. Instead of detailed modelling, requiring discrete event simulations and detailed specification of control mechanisms under different contingent events, we use measurements from transactional systems that provide information about the effect of human interventions, leading to measurement of intervention-independent performance and intervention-dependent performance, the latter being typically higher. Further research is needed to develop this approach in a generic modelling approach, capable of dealing with causal chains of events that create mutual dependency between relevant stochastic processes.

7 References

Axsäter, S. (1990). Modelling emergency lateral transshipments in inventory systems. *Management Science, 36*(11), 1329-1338.

Axsäter, S. (2003). A new decision rule for lateral transshipments in inventory systems. *Management Science, 49*(9), 1168-1179.

Barad, M. (2013). Flexibility development–a personal retrospective. *International Journal of Production Research*, 51(23-24), 6803-6816.

Barad, M., & Sipper, D. (1988). Flexibility in manufacturing systems: definitions and Petri net modelling. *International Journal Of Production Research*, 26(2), 237-248.

Beach, R., Muhlemann, A. P., Price, D. H. R., Paterson, A., & Sharp, J. A. (2000). A review of manufacturing flexibility. *European journal of operational research, 122*(1), 41-57.

Bertrand, J. W. M. (2003). Supply chain design: flexibility considerations. *Handbooks in Operations Research and Management Science*, *11*, 133-198.

Bertrand, J. W. M., Wortmann, J. C., & Wijngaard, J. (1990). Production control: a structural and design oriented approach.

Dada, M., (1992). A Two-Echelon Inventory System with Priority Shipments. *Management Science* 38, 1140-1153.

De Kok, A.G. (2011). Inventory control with manufacturing lead time flexibility. BETA working paper series, 345. Eindhoven: Technische Universiteit Eindhoven, 20 pp.

De Kok, T.G. (2015). Buffering against uncertainty in high-tech supply chains. In *Winter Simulation Conference (WSC), 2015* (pp. 1-10). IEEE.

De Kok, T.G., & Fransoo, J. C. (2003). Planning supply chain operations: definition and comparison of planning concepts. *Handbooks in operations research and management science*, *11*, 597-675.

De Toni, A., & Tonchia, S. (1998). Manufacturing flexibility: a literature review. *International journal of production research*, *36*(6), 1587-1617.

Fransoo, J.C., V. Sridharan, and J.W.M. Bertrand (1995). A hierarchical approach for capacity coordination in multiple products single-machine production systems with stationary stochastic demands. *European Journal of Operational Research* 86, 57-72.

Hadley, G. and T.M. Whitin (1963). *Analysis of Inventory Systems*. Prentice-Hall, Englewoods Cliffs.

Inderfurth, K., & Minner, S. (1998). Safety stocks in multi-stage inventory systems under different service measures. *European Journal of Operational Research*, *106*(1), 57-73.

Jain, A., Jain, P. K., Chan, F. T., & Singh, S. (2013). A review on manufacturing flexibility. *International Journal of Production Research*, *51*(19), 5946-5970.

Janssen, F. B. S. L. P., Heuts, R., & De Kok, T.G. (1999). The impact of data collection on fill rate performance in the (R, s, Q) inventory model. *Journal of the Operational Research Society*, 50(1), 75-84.

Lee, H.L. and S. Nahmias (1993). Single-Product, Single-Location Models. in: Logistics of Production and Inventory, edited by Graves, S.C., Rinnooy Kan, A.H.G. and Zipkin, P.H., North-Holland, Amsterdam.

Minner, S., Diks, E. B., & De Kok, A. G. (2003). A two-echelon inventory system with supply lead time flexibility. *IIE Transactions*, *35*(2), 117-129.

Olhager, J. (1993). Manufacturing flexibility and profitability. *International journal of production economics*, 30, 67-78.

Romme, J., & Hoekstra, S. (Eds.). (1992). *Integral Logistic Structures: Developing Customer-oriented Goods Flow*. Industrial Press.

Seebacher, G., & Winkler, H. (2013). A citation analysis of the research on manufacturing and supply chain flexibility. *International Journal of Production Research*, *51*(11), 3415-3427.

Silver, E. A., D. F. Pyke, and R. Peterson (1998). *Inventory Management and Production Planning and Scheduling*, 3rd Edition, John Wiley & Sons, New York.

Simpson Jr, K. F. (1958). In-process inventories. *Operations Research*, *6*(6), 863-873.

Smits, S. R., Wagner, M., & de Kok, T. G. (2004). Determination of an order-up-to policy in the stochastic economic lot scheduling model. *International Journal of Production Economics*, *90*(3), 377-389.

Van Houtum, G. J., & Kranenburg, B. (2015). *Spare parts inventory control under system availability constraints.* International Series in Operations Research & Management Science, 227. Springer.

Verrijdt, J. H. C. M. (1997). *Design and control of service part distribution systems*. Technische Universiteit Eindhoven. Unpublished PhD thesis.

Verrijdt, J.H.C.M., I.J.B.F. Adan, and A.G. de Kok (1998). A tradeoff between emergency repair and inventory investment. *IIE Transactions* 30(2), 119-132.

Vollmann, T.E., W.L. Berry, D.C. Whybark, and F.R. Jacobs (2004). *Manufacturing Planning and Control for Supply Chain Management*, McGraw-Hill/Irwin.

Zhang, Q., & Tseng, M. M. (2009). Modelling and integration of customer flexibility in the order commitment process for high mix low volume production. *International Journal of Production Research*, *47*(22), 6397-6416.