


---

**Monograph:**
Riddalls, C.E. and Bennett, S. (2000) Production-Inventory System Controller Design and Supply Chain Dynamics. Research Report. ACSE Research Report 769 . Department of Automatic Control and Systems Engineering

---



# Production-Inventory System Controller Design and Supply Chain Dynamics

Research Report No 769

C.E. Riddalls and S. Bennett

Department of Automatic Control and Systems Engineering, University of Sheffield,

S1 3JD. Tel 0114 222 5186, fax 0114 273 1729, c.e.riddalls@sheffield.ac.uk

ABSTRACT: *This paper deals with the modelling and control of aggregated production-inventory systems as described by differential equations. Hitherto, research in this area has been characterised by the approximation of production delays by first order lags, rather than more realistic pure delays. We demonstrate the substantial qualitative differences between these two approaches and thus generate the motivation for the rest of the paper, which tackles pure delay systems. The application of some relatively new design methodologies for delay systems yields four design choices, which are tested for their performance over a range of criteria including stability robustness. This investigation is then extended to the model of a supply chain comprising many such production-inventory systems. The mechanism by which disturbances can be transmitted along the supply chain causing disruption and incurring costs to other supply chain echelons is elucidated. An heuristic feedback policy designed to adaptively tune the individual system designs in response to such disturbances is presented.*

## 1. Introduction

This paper deals with the modelling and control of aggregated production-inventory systems as described by differential equations. Hitherto, research in this area has been characterised by the approximation of production delays by first order lags, rather than more realistic pure delays. By example, (see section 2) we demonstrate the substantial qualitative differences between these two approaches and thus generate the motivation for the rest of the paper. The novelty of our approach derives from the treatment of pure delays which produce much more accurate models of production-inventory systems. The application of some relatively new design methodologies for delay systems yields four design choices, which are tested for their performance over a range of criteria (section 3). The inclusion of pure time delays enables the investigation of their influence on system stability. This is an important question since logistical disturbances to such systems are usually manifested in increased delays somewhere along the production process which may have unforeseen ramifications for the dynamics of the entire supply chain. So in section 4 this investigation is extended to the model of a supply chain comprising many such production-inventory systems. The mechanism by which disturbances can be transmitted along the supply chain causing disruption and incurring costs to other supply chain echelons is elucidated. Understanding the pathology of such phenomena enables the development of policies for their elimination or mitigation. Indeed, in the last section an heuristic feedback policy designed to adaptively tune the individual system designs in response to such disturbances is presented.

## 2. Modelling Production-Inventory Systems

For fifty years [Simon, 1952] control theorists have been attempting to apply their discipline to the modelling and control of production-inventory systems. Yet the widespread adoption in industry of production control techniques developed in the field of Operational Research [Naddor 1966, Elsayed & Boucher, 1985] has subordinated these efforts. Consequently, tools from control theory have, more recently, been confined to simulation and the development of generic strategic recommendations [Wikner et al. 1991, Towill, 1991], rather than as a practical aid to decision making. In this section we briefly review the relevant literature and highlight the common deficiency of these methods, the substitution of exponential smoothing delay approximations for pure delays. This, we claim, constitutes the most significant obstacle to their greater acceptance and serves as a motivation for the development of the rest of the paper

Control-theoretic methods in this area fall into two qualitatively quite distinct categories. These are cost based and non-cost based approaches. The former use optimal control to trade off the costs of holding inventory against production costs [Bensoussan and Proth 1982, Lieber 1973]. By their very nature they are a planning tool and so require a demand forecast to generate a production plan. In contrast, the methods considered in this paper are designed to react in real time to varying demands. They attempt to arrive at production strategies which reject ephemeral demand variations whilst satisfying substantive demand patterns through policies which avoid both volatile production rates and prolonged excessive inventory safety stock depletion. Typically, by tuning parameters in a simple state feedback representation of the system, its transient behaviour in response to a likely demand variation is shaped to partially satisfy these conflicting requirements. These methods possess the advantage over those using optimal control that they require no detailed

knowledge of the (often complicated) production and inventory holding cost structures. Further, knowledge of the particular demand patterns involved can be used to fine tune the specific system design. For instance, if it is known that demand fluctuations in a certain frequency range tend to occur regularly, then the design can focus on shaping the frequency response of the system over that range.

Every differential equation model of a production-inventory system must be constructed around the fundamental inventory balance equations:

$$\frac{di}{dt} = p^c(t) - d(t) \tag{1}$$

where $i(t)$ is the inventory level, $p^c(t)$ is the production completion rate and $d(t)$ is the demand, rate all at time $t$. Using simple Laplace transform feedback techniques to derive rational controllers, Simon [1952] designed systems that aimed to keep the inventory near a desired level whilst minimising production fluctuations. This early paper highlighted a fact which persists in influencing research in this area today, namely that pure delays are hard to deal with within a differential equation framework. In common with many other authors, to circumvent the analytical intractability caused by pure delays, he approximated them by distributed lags (also called exponential smoothing), i.e. another first order differential linear differential equation yielding an exponential solution. This approach was justified by regarding the Laplace domain lag approximation as the transform of a probability density function which expressed the probability that the lag in producing any particular item would be a particular value. However, Figure 1 shows how first order exponential lags exhibit qualitatively very different behaviour to pure time delays in response to likely demands placed upon the production facility. These demands are given by the

function $p_d(t)$, standing for demanded production. In the Laplace domain the equations governing the processes are

$$\frac{P^c(s)}{P^d(s)} = \frac{1}{1+Ts}, \quad \frac{P^c(s)}{P^d(s)} = e^{-Ts}, \tag{2}$$

representing a first order lag with parameter $T$ and the equivalent pure time delay, respectively (Laplace transforms being denoted by upper case letters).

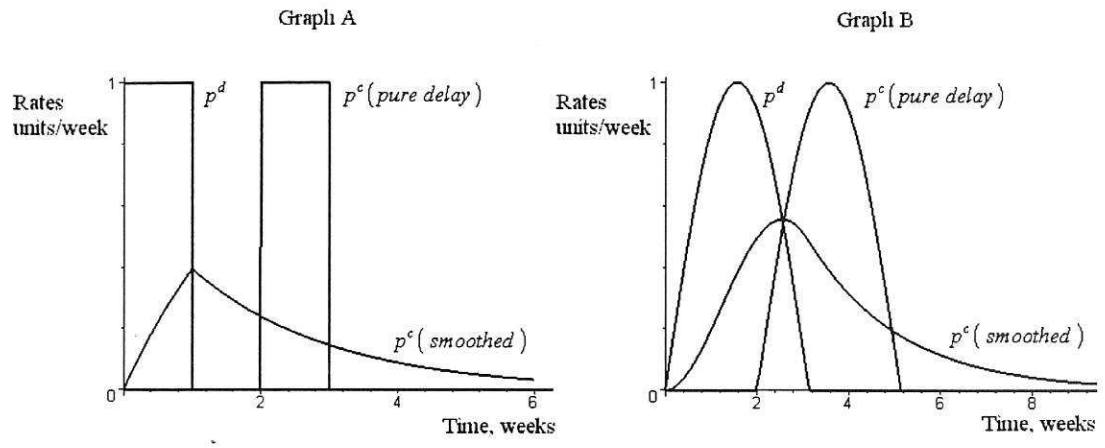


Fig. 1. Pure and exponentially smoothed delays.

Graphs A and B show the responses to a transient unit step and to a sinusoidal-type blip, respectively. Notice, in particular, the differing attenuating effect of the exponential smoothing process on $p^c(t)$(smoothed) in response to the qualitatively different signals. It is possible to vary the parameter $T$ to achieve different responses: Increasing $T$ smooths out $p^c(t)$ at the expense of more attenuation, whilst decreasing $T$ results in responses which look more like $p^d(t)$. This means that first order lags can approximate very small delays with only slight attenuation, however the more accurate representation of pure delays of small magnitude can be dealt with using classical design techniques [Marshall, 1979]. Marshall quantifies what a significantly

'large' pure delay is, beyond which these methods become insufficient. If the design bandwidth is $\omega$, then $T$ will be 'large' if $\omega T$ is of the order $\pi/2$. So, even if $T$ is small, a large design bandwidth precludes the use of classical techniques. Moreover, for all the controllers used in this paper, on systems which were chosen for their plausibility, $\omega T > \pi/2$ (see section 3).

Some authors have used higher order distributed lags to approximate pure delays [MacDonald, 1978]. These increase the dimension of the system, inhibiting analytical progress, and merely lead to sums of terms similar to that in $p^c(t)$(smoothed) above, increasing the resemblance to $p^d(t)$ only marginally.

Despite these drawbacks, the recondite nature of delay differential equations has discouraged their application to this area and instead many authors have modelled production-inventory systems using distributed delays. Forrester [1961] developed detailed nonlinear models of multi-echelon production-inventory systems. The complexity of these models precluded detailed analysis and instead they were used simply for simulation purposes. Unfortunately any system design then amounted to a process of trial and error and so the utility of such models was restricted to the qualitative investigation of 'what if' scenarios. Their complexity also discouraged attempts at their rigorous validation. By performing an elementary sensitivity analysis on his model, Forrester would have discovered a high degree of parameter sensitivity in the exponential delays. In today's parlance, this lack of robustness was exacerbated by the repeated coupling of similar submodels. Forrester was duly criticised for the lack of theoretical underpinning in these models (Ansoff and Slevin 1968) and later for similar work modelling world populations (Forrester 1973). Given the limited analytical utility of such models and their resulting classification as simulation

models, the many sophisticated discrete event simulation packages available today may provide a more accurate simulation capability.

Acknowledging these issues, Towill [1982] simplified Forrester's models (in this paper, in single-echelon form) to an extent at which they were amenable to analytic study. He then used classical control techniques to tune the parameters in the feedback (of the inventory level) and feedforward (of the demand rate) loops. His work shows qualitatively how the various design objectives can be traded off against one another. However, the use of distributed lags must still compromise the fidelity of the model and consequently our faith in the design rules. Also, no analysis linking the system's stability to the production time constant is presented, an important omission since logistical disturbances tend to increase the production delays, possibly leading to instability. So the design should be robust for a range of delays about the nominal value. In [Porter and Taylor, 1972] the authors use similar feedback techniques to achieve a desirable transient response of the system. In [Bradshaw and Porter, 1975] this work is extended to allow for the stimulation of sales by advertising.

Pure time delay systems frustrate significant analytical progress because, in the Laplace domain, the most commonly chosen design arena for engineers, they yield transcendental characteristic equations with (in general) infinitely many roots. As for linear systems, the behaviour of the system is determined by these roots [Bellman and Cooke, 1963], but, since there are infinitely many of them, the usual design methodology of shifting them to more desirable locations is nontrivial. To the authors' knowledge, the only paper to deal with pure delay production-inventory system is [Mak et al., 1976]. In this paper the feedback parameter values which yield stable systems are calculated given fixed pure delays. This process is repeated for different delay values to give a qualitative picture of the overall stability

characteristics of the system. It is shown that as the delay increases the range of feedback parameters resulting in a stable system is reduced. However, explicit delay-dependent stability regions are not calculated and no design methodology for determining the feedback parameters is given. As an aside, we are puzzled at the inclusion of an exponential smoothing term as well as a pure time delay in their model.

## 3. The design of Production-Inventory systems with pure delays

*Design Objectives*

According to Towill [1982], the two main design objectives for production-inventory systems are good inventory recovery from deterministic demand changes and good rejection of random demand disturbances. Deterministic demand patterns are those whose magnitude and sustained nature make it most profitable to satisfy by changing the production rate and (possibly) depleting safety stocks, but avoiding the risk of stockouts. In contrast, random demand disturbances are those which are considered either too small or too ephemeral to warrant a costly radical change in production and therefore should be met by safety stock depletion and possible transient toleration of stockouts. The particular cost structure of the system determines where the boundary between such demand patterns lies. Further, this boundary may change over time, depending on the capacity utilisation of the production process. For example, for some production processes it is very expensive to change the production rate substantially due to increased setup times, machine calibration changes, etc. For these systems, small and irregular fluctuations in demand should mainly be satisfied by a buffer stock, whose cost of maintenance should be reconciled against the exorbitant cost of significant production variations. However, more fundamental demand

variations, over longer periods cannot be satisfied by safety stock depletion alone and so the production rate must change in order to avoid prolonged periods of stockouts. A conflict between the two objectives of following deterministic demands and ejecting random demand disturbances arises since improving the recovery rate of the system to dterministic demands invariably makes the system more responsive to all demand variations. However, by using the fact that random demand disturbances tend to be of a higher frequency than deterministic demands, we can shape the frequency response of the system to attenuate high frequency signals and follow low frequency signals.

We will add the further design objective: To ensure the system is stable for a range of time delays about the nominal value. This is important since logistical disturbances to the system (e.g. late deliveries, quality problems, machine breakdowns, absent workers) can be manifested in increased production times. Elementary systems theory tells us that long delays can lead to instability, which translates into erratic production rates, an expensive logistical phenomenon. Further, this instability can be communicated along the supply chain, resulting in the well-known bullwhip effect. Also known as the demand amplification effect, this is the tendency of demand variations to be amplified as they are passed down the supply chain [Houlihan 1987, Burbidge 1984]. Working with this wider supply chain perspective promotes the objective of robust design methodologies that ensure stability for a range of delays. We shall examine the supply chain ramifications of production-inventory system design in more detail in section four.

Figure 2 shows the structure of our production-inventory system. The infinite dimensional nature of delay systems implies that classical transfer function (and thus block diagram) representations of these systems are inherently ambiguous [Fliess and

Mounier, 1998]. So we simply regard figure 2 as a description of the information flows in the system.

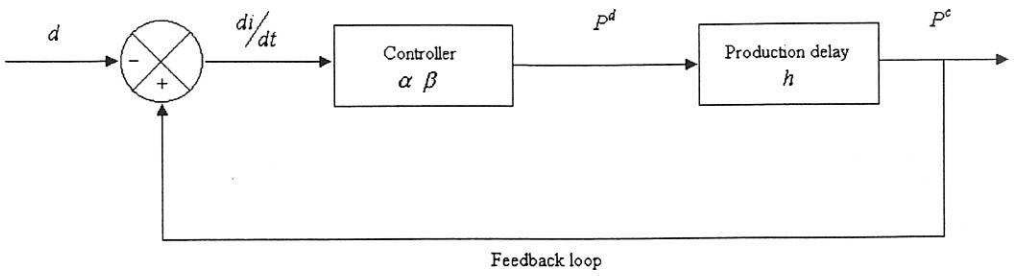


Figure 2. Production-Inventory control system

Once again, $i(t)$ is the inventory level, $p^c(t)$ is the production completion rate, $p^d(t)$ is the demanded production rate (i.e. the scheduled production rate) and $d(t)$ is the demand rate, all at time $t$. The parameter $h$ is the length of the fixed pure time delay comprising the time taken to process the order, take delivery of raw materials (if necessary) and manufacture the product. We assume that over the useful time horizon of the model the demand varies about a constant level $\bar{d}$. With the system in equilibrium $d(t) = p^c(t) = p^d(t) = \bar{d}$. Hence we can take this constant value off each of the listed rate quantities to simplify the equations of the system. These are

$$\frac{di}{dt} = p^c(t) - d(t), \quad p^c(t) = p^d(t-h) \tag{3}$$

We aim to keep the inventory level close to a desired value (the safety stock level) which, without loss of generality, we can take to be zero. Hence $i(t)$ really denotes the inventory level discrepancy (from zero). This approach treats stockouts exactly the same as inventory depletion. The most simple form of controller (as in [Towill, 1982]) monitors the inventory discrepancy and schedules future production to be proportional to it. Hence

$$p^d(t) = \alpha i(t),\ \alpha < 0 \tag{4}$$

We shall also examine the performance of the feedback rule

$$p^d(t) = \alpha i(t) + \beta \frac{di}{dt},\ \beta < 0 \tag{5}$$

which, through the inclusion of the derivative term, can compensate changes in the inventory level earlier and also increase the ability to smooth the demands placed upon the production facility. In classical engineering terms, if $\frac{di}{dt}$ is the state of the system, then (5) constitutes proportional plus integral control action, a favoured design choice [Gorecki et al., 1989]. Putting (3), (4) and (5) together yields

$$\frac{di}{dt} = \alpha i(t-h) - d(t) \tag{6}$$

$$\frac{di}{dt} = \alpha i(t-h) + \beta \frac{di}{dt}(t-h) - d(t) \tag{7}$$

for the closed loop equations of the two systems. Many authors include either a demand feedforward loop [Towill, 1982] or an extra feedback term comprising the integral of the inventory discrepancy [Porter and Taylor, 1972; Bradshaw and Porter, 1975]. Both these loops have the effect of eliminating the inventory steady state error when $d(t) \to d \neq 0$, as $t \to \infty$. Our approach presupposes the constancy of an average demand rate (taken to be zero, without loss of generality) over a planning horizon which will depend on the demand characteristics, obviating the need for these extra loops. We have restricted our remit to the design of systems that react in a desired manner to perturbations in demand about an average steady state level. There are sound reasons for treating these issues separately from the steady state system design. The latter favours high volumes for which the unit cost of production is paramount and flexibility is subordinated, the former comes into play during periods of turmoil in the supply chain: promotional periods, periods during which there are

logistical disturbances to the system. Thus these designs favour responsiveness and swift recovery. The timescales to be considered in the design of the steady state system (e.g. those determined by the product lifecycle or seasonal variations) are much longer than those in the transient system design (e.g. those determined by promotions, machine breakdowns, heatwaves etc). Fundamental changes to the steady state system-like long term changes in the average demand rate-also have cost implications best analysed by the structural tools of Operational Research [Naddor, 1966] rather than the tactical decision making tools presented here. For instance, an increase in this rate should prompt a re-evaluation of the safety stock level, thus affecting the whole cost structure of the system. Furthermore, the periodic review of demand characteristics, demanded by our approach, with the intention of fine tuning the choice of $\alpha$ and $\beta$, should lead to better performance.

*Design Methodology*

Many linear finite-dimensional system design methodologies consist of adjusting the feedback parameters so that the roots of the characteristic equation have minimum real parts. This leads to very responsive systems with quickly decaying transients, but, if used without care, can also result in responses which are too volatile. The extension of these techniques to time delay systems must proceed with circumspection since there are infinitely many characteristic roots to consider. Stabilising some of them might increase the complex parts of others, leading to more oscillation. We shall use a generalisation of the aperiodic stability criterion as presented in the book [Gorecki et al., 1989]. As applied to finite-dimensional systems, the regulator is chosen so that there exists a real characteristic root of maximal attainable multiplicity. Unlike the case of finite-dimensional systems, this does not deliver guaranteed non-oscillatory transients. However, we can expect that, for small delays, the transient components

corresponding to complex characteristic roots will decay much quicker than those corresponding to the real characteristic root. In any case, for all the design procedures presented here we shall check the resulting stability margins of the closed loop system and qualitatively investigate the response of the system to some typical demands.

By taking the Laplace transform of (6) and rearranging, we obtain the characteristic equation for the transfer function of $P^c/D$:

$$g(s) = \alpha e^{-hs} - s \tag{8}$$

If $g(s)$ has a root, $r$, of multiplicity $k$, then $g^{(j)}(r) = 0,\ j = 0 \ldots k-1,\ g^{(k)}(r) \neq 0$.

Repeatedly differentiating (8), we get

$$\dot{g}(r) = -h\alpha e^{-hs} - 1,\quad \ddot{g}(r) = h^2 \alpha e^{-hs} \tag{9}$$

We see that $\dot{g} \neq \ddot{g}$, hence the maximal order of the characteristic root is two. Now, setting $g = \dot{g} = 0,$ and solving, we have

$$r = \frac{-1}{h},\quad \alpha = \frac{-e^{-1}}{h} \tag{10}$$

We shall call this design 1. For the case when $\beta \neq 0$ the extra degree of freedom in the design generates a third order root from the simultaneous equations

$$\begin{aligned} g(r) &= (\alpha + \beta s)e^{-hs} - s = 0 \\ \dot{g}(r) &= (\beta - \alpha h - \beta h s)e^{-hs} - 1 = 0 \\ \ddot{g}(r) &= (-2\beta h + \alpha h^2 + \beta h^2 s)e^{-hs} = 0 \end{aligned} \tag{11}$$

yielding,

$$r = \frac{-2}{h},\ \alpha = \frac{-4e^{-2}}{h},\ \beta = -e^{-2} \tag{12}$$

These parameters constitute design 2. Notice that the smoothing parameter $\beta$ is independent of the system delay. This is not surprising since it acts on the derivative rather than the absolute value of the inventory level. We shall examine the

performance of these parameter values once we have presented the third and fourth designs.

Production-inventory system design in the frequency domain is particularly attractive because the knowledge and experience of the production manger can be utilised to focus attention on specific bands in the frequency range. For instance, looking at past data might highlight recurrent demand oscillations at certain frequencies. The frequency response of the system can then be shaped in the vicinity of these frequencies to be flat so that these demands are reproduced accurately. In contrast, ephemeral disturbances in demand can be rejected by attenuating the high frequency range of the response. Unfortunately, the transcendental nature of the characteristic equation makes the simultaneous achievement of both of these objectives problematic. One method [Marsik, 1958; Górecki et al., 1989] attempts to make the low-frequency part of the attenuation diagram flat and relies on the existing high frequency attenuation of most controllers to suppress spurious disturbances. We accomplish this by choosing the regulator settings so that

$$\frac{d^k}{d\omega^k}\left|G(j\omega)\right|^2_{\omega=0} = 0,\ k = 1,\ldots,l \tag{13}$$

where $G$ is the transfer function of $P^d/D$ and $l$ is chosen as large as possible. Put $\gamma(\omega) = |G(j\omega)|^2 = L(\omega)/M(\omega)$, where $L$ and $M$ real symmetric functions. Since all odd derivatives of $\gamma(\omega)$ are zero, we can take $l$ to be even. Then

$$L(0)M^{(2k)}(0) = M(0)L^{(2k)}(0),\ k = 1,\ldots,l/2 \tag{14}$$

The number of adjustable parameters determines $l/2$. For the system (6)

$$G(j\omega) = \frac{\alpha}{\alpha e^{-j\omega h} - j\omega}, \tag{15}$$
$$M(\omega) = \alpha^2 + \omega^2 + 2\alpha\omega\sin\omega h,\ L(\omega) = \alpha^2$$

and

$$M^{(2)}(0) = 2 + 4\alpha h \cos \omega h - 2\alpha h^2 \sin \omega h\Big|_{\omega=0} = 0 \tag{16}$$

giving

$$\alpha = \frac{-1}{2h} \tag{17}$$

Similarly, for (7), we have

$$\begin{gathered} M(\omega) = \alpha^2 + (1+\beta^2)\omega^2 + 2\alpha\omega \sin \omega h - 2\beta\omega^2 \cos \omega h, \\ L(\omega) = \alpha^2 + \beta^2\omega^2 \end{gathered} \tag{18}$$

Hence (14) yields

$$\begin{gathered} M(0) = L(0) = \alpha^2 \\ M^{(2)}(0) = 2 + 2\beta^2 + 4\alpha h - 4\beta = 0,\ L^{(2)}(0) = 2\beta^2 \\ M^{(4)}(0) = 24\beta h^2 - 8\alpha h^3 = 0,\ L^{(4)}(0) = 0 \end{gathered} \tag{19}$$

Solve to give

$$\alpha = -3/4h,\quad \beta = -1/4 \tag{20}$$

These last two designs will be referred to as designs 3 and 4, respectively, in the order in which we derived them.

*Stability*

In the appendix we derive the stability regions for a cascaded production-inventory system. So as not to repeat ourselves, we quote here the results from that section for single echelon systems. A technical definition of asymptotic stability can be found in [Bellman and Cooke, 1963] but the reader may simply make do with the idea that stability is the property which brings the system back into equilibrium after a perturbation in demand. Conversely, the production rate of an unstable system will grow indefinitely with increasing time. Assuming $\alpha < 0$ and $\beta < 0$, the systems (6) and (7) are asymptotically stable if

$$0 < h < \frac{-\pi}{2\alpha} \; ; \; 0 < h < \frac{-\sqrt{1-\beta^2}}{\alpha}\cos^{-1}\beta \text{ and } |\beta| < 1 \qquad (21)$$

respectively. Figure 3 shows the nature of this stability region.

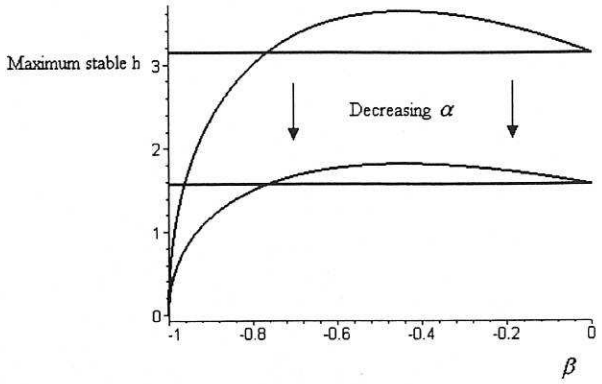


Figure 3. Stability regions for a production-inventory system

The curved lines plot the maximum value of $h$ for the system (7) to be stable whilst the straight lines (independent of the value of $\beta$) do so for (6). The lower curves are for $\alpha = -1$ and the higher curves for $\alpha = -0.5$. We can see that for all but the smallest values of $\beta$, the system (7) is more robust than (6) since it remains stable for a bigger range of delay. However, decreasing $\alpha$ cuts this advantage until it's almost negligible. This diagram shows that attempting to stabilise the system by decreasing $\alpha$ impairs its robustness and, paradoxically, can lead to instability. As we have said, this phenomenon is not reproduced by models which use exponential smoothing to mimic delays.

We now calculate the specific $h-$stability margins $(V)$ for the for designs derived in the last section. We define this to be the maximum delay for which the system is stable, and express it as a multiple of the nominal delay value, $h$. Substituting the relevant values for $\alpha$ and $\beta$ (where appropriate) into (21) yields

| **Design** | $\alpha$ , $\beta$ | $h-$**stability margin,** $V$ |
|---|---|---|

| | | |
|---|---|---|
| **1** | $\alpha = -e^{-1}/h, \beta = 0$ | $4.27h$ |
| **2** | $\alpha = -4e^{-2}/h, \beta = -e^{-2}$ | $3.12h$ |
| **3** | $\alpha = -1/2h, \beta = 0$ | $3.14h$ |
| **4** | $\alpha = -3/4h, \beta = -1/4$ | $2.35h$ |

Table 1. System designs and stability margins

*System Performance*

In this section we evaluate the performance of the four controllers in response to likely demands placed upon them. Engineers typically achieve this by applying a unit impulse or step function to the system [Ogata, 1990]. However, unit impulses are not encountered in real life and so should be avoided. Further, since we have assumed that the average demand rate over a significant length of time is zero, it would be better to use a transient step rather than a pure step in demand as a typical input to the system. Sinusoidal-type demands and transient step demands are, we believe, much more prevalent in real situations. Our aim is to evaluate the particular balance for each design between noise rejection, fast inventory depletion recovery and avoidance of oscillatory or volatile production patterns.

Figure 4 shows the behaviour of the systems with $h = 4$ days in response to a transient step in demand. To save space, we have presented rates and levels on the same graph with different scales for each. It is important to remember that these graphs represent fluctuations about an average level which, for simplicity, we have taken to be zero. The upper left graph is labelled, with the others following the same pattern. Figure 5 shows the behaviour of the four systems with $h = 3$ days in response to a noise-corrupted step in demand. The sinusoidal noise element in the demand pattern in figure 5 was included to facilitate the investigation of the demand noise suppression

of each design. For this reason the simulations where continued after the step recovery period to demonstrate the steady state production rates in response to persistent perturbations. The substantive component of the demand signal was chosen to be a transient step, so as to highlight the differences with the sinusoidal demand. The noise is modelled by a high frequency sine wave, the amplitude of which was chosen to be a significant 20% of the substantive step change. Its frequency was chosen to mimic daily fluctuations in orders about the step change. A higher frequency choice would be unrealistic since, for most manufacturers, changes in demand during one day are not monitored, but, rather, aggregated into an order at one point during that day. However, for some applications, higher frequency noise signals might be worth examining. For example, supermarkets monitor demand throughout the day and, for some products, may even formulate ordering policies based on this real time data.

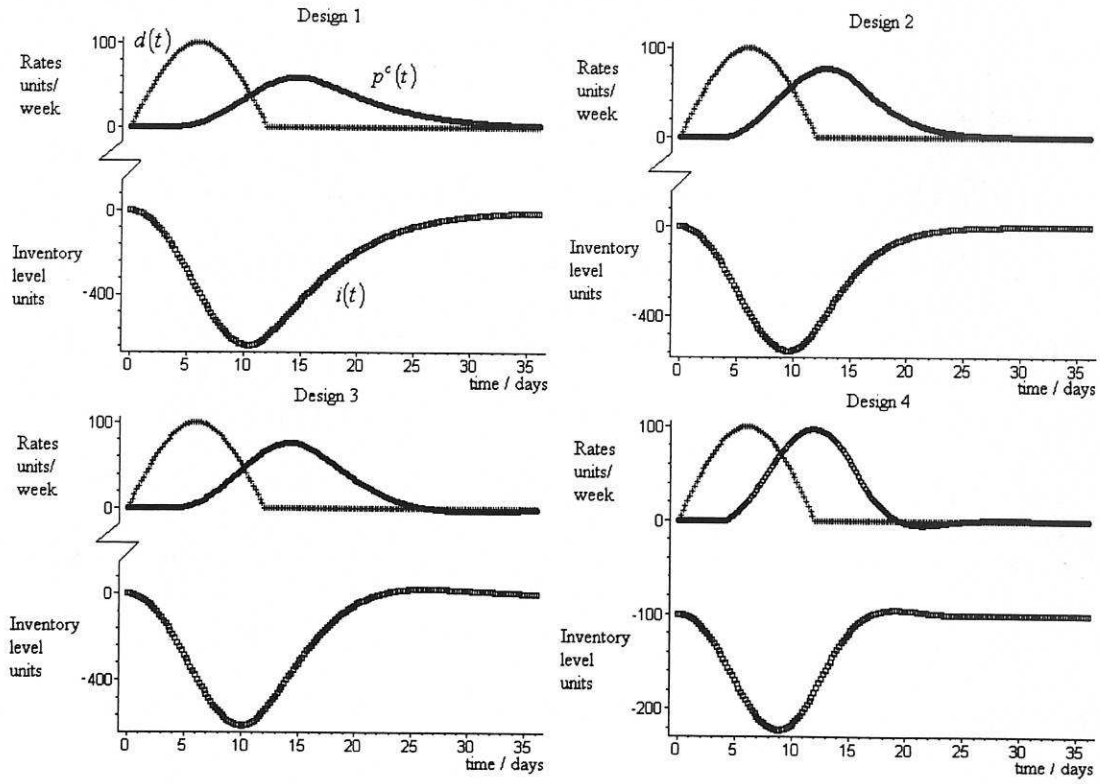


Figure 4. Response of system to sinusoidal-type demand, $h = 4$.

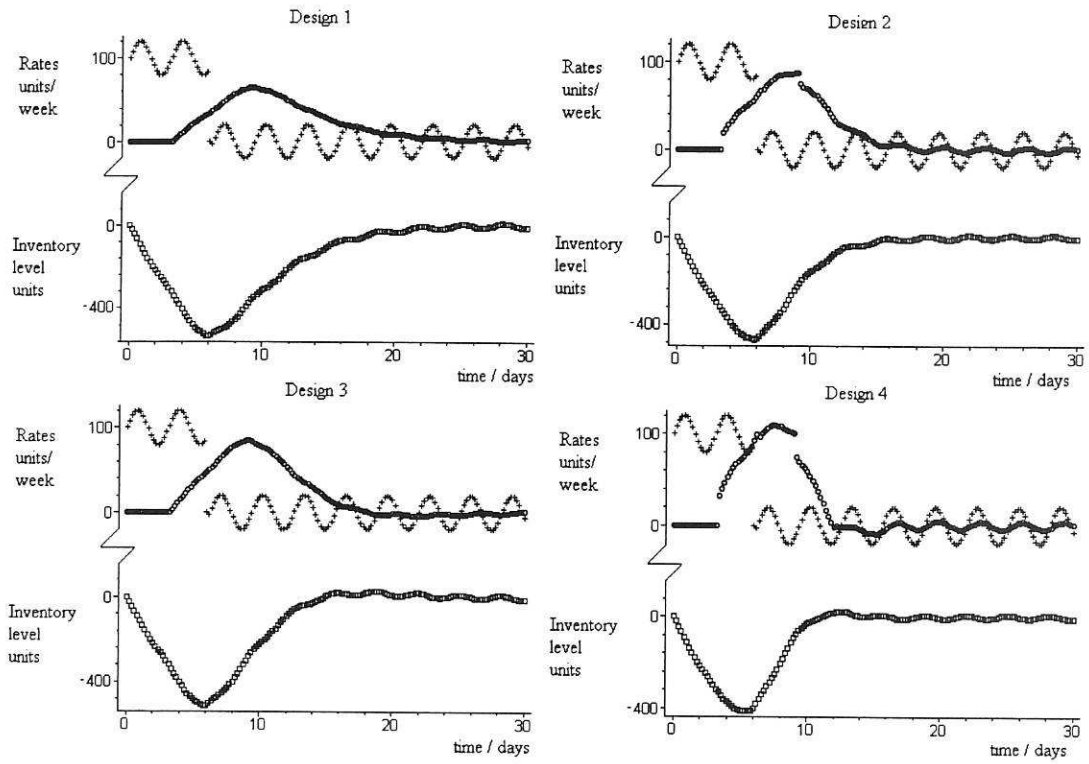


Figure 5. Response of systems to noise-corrupted transient step in demand, $h = 3$.

The qualitative results of these experiments are presented in table 2, where we have ranked the performance of each design for each characteristic (best at the top, worst at the bottom).

| Robust Stability | Smooth Production | | Inventory Recovery | | Noise Rejection |
|---|---|---|---|---|---|
| | $h = 3$ | $h = 4$ | $h = 3$ | $h = 4$ | |
| 1 | 1 | 1 | 4 | 4 | 1 |
| 3 | 3 | 3 | 3 | 2 | 3 |
| 2 | 2 | 2 | 2 | 3 | 2 |
| 4 | 4 | 4 | 1 | 1 | 4 |

Table 2. Table showing performance of various control designs.

All of the designs considered yield trajectories which exhibit little or no low frequency (production) oscillation once the initial perturbation is corrected. For this reason we have omitted a column for this characteristic. In general terms, we see that there is an obvious trade off between smooth production and swift inventory recovery. For $h = 4$ and $h = 3$ we see that design 1 has the smoothest response and inevitably the slowest inventory recovery rate. Design 4, in contrast, produces the most responsive production pattern which closely matches the demand pattern (after the given delay). This is achieved at the expense of a very slight overshoot in the inventory correction and the transmission of the demand discontinuity to the production pattern. Designs 2 and 3 exhibit very similar behaviour, being a compromise of the two extremes of design 1 and 4. The knowledge and expertise of the production manager should be heeded when choosing a design. As expected, the most responsive system (design 4) suppressed the high frequency demand signal the least but, even for this design, the amplitude of the steady state production rate oscillation is substantially less than that of the demand signal and so some attenuation is achieved. For the other designs most of these fluctuations are absorbed by inventory, rather than production changes (a lesser evil). There is also a direct correlation between stability robustness and noise rejection at the given frequency of noise. It is important to note that, again these results only tell us about the behaviour of the system in response to noise at that given frequency. In practice, the system design might need tuning to meet the demands of a specific collection of typical demands. In any case, the designs we have presented may serve as a starting point for this fine tuning by varying $\alpha$ and $\beta$. Decreasing $\alpha$ improves the responsiveness of the system but may lead to oscillation and instability. Decreasing $\beta$ smooths the system response, increasing the inventory depletion correction time and damping

oscillation. However, too small a value of $\beta$ can lead to oscillation and instability (see (21)). The stability properties of each design should always be checked.

## 4. The Dynamics of Supply Chains

Production-inventory systems rarely exist in isolation, rather, they connect together in series and parallel to form an, often complex, supply chain. Practitioners in this field have known for some time that the design of production-inventory systems in isolation, without reference to the rest of the supply chain can lead to poor performance in the overall supply chain. One of the best known manifestations of this poor performance is the demand amplification effect [Towill 1992, Houlihan 1987, Burbidge 1984], the tendency of demand fluctuations to be amplified as they are communicated down the supply chain. Using cascaded series of the familiar 'exponential smoothing' models (see section 2), a series of papers has investigated this phenomenon [Wikner et al. 1991, Towill, 1991, Towill, 1992, Berry et al. 1995, Towill and Del Vecchio, 1994]. In general terms, they conclude that demand amplification can be suppressed by shortening lead times, an intuitively reasonable conclusion despite the criticisms of exponential smoothing made in section 2 which can also be levelled at these models. In this section we construct a supply chain by cascading the systems designed in section 3, and show that the robustness of any particular echelon can affect the performance of the rest of the chain. We then demonstrate how logistical disturbances can cause demand amplification and then present a heuristic feedback methodology to counter these effects.

Consider the generalisation of figure 2 to the $N-$echelon model shown in figure 6.

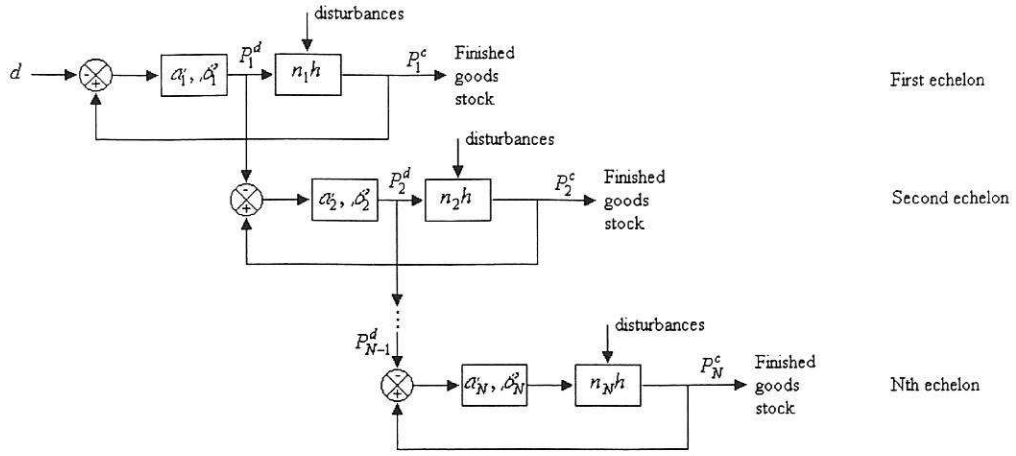


Figure 6. $N$ −echelon supply chain.

The production delays are assumed to be integer multiples of some small number $h$. This representation facilitates considerably the calculation of the stability properties of the system (see appendix) which, we consider, would be intractable given the definition of arbitrary delays. In fact, for real systems, there will invariably exist a minimum time period (e.g. 1 day or ½ day) of which all other delays are an integer multiple. The system equations are

$$\begin{aligned}
\frac{di_1}{dt} &= \alpha i_1(t-n_1h)+\beta_1\frac{di_1}{dt}(t-n_1h)-d(t)\\
\frac{di_2}{dt} &= \alpha i_2(t-n_2h)+\beta_2\frac{di_2}{dt}(t-n_2h)-\left\{\alpha_1 i_1(t)+\beta_1\frac{di_1}{dt}(t)\right\}\\
&\vdots\\
\frac{di_N}{dt} &= \alpha i_N(t-n_Nh)+\beta_N\frac{di_N}{dt}(t-n_Nh)-\left\{\alpha_{N-1} i_{N-1}(t)+\beta_{N-1}\frac{di_{N-1}}{dt}(t)\right\}
\end{aligned} \tag{22}$$

So the demands placed upon the *ith* echelon are taken to be proportional to the demanded production rate of the $(i-1)st$ echelon ($P_{i-1}^d$), as in all the simulation models quoted above. For $\alpha<0,-1<\beta<0$, this system is stable for $0<h<h^*$, where

$$h^* = \min_i V_i = \min_i \frac{-\sqrt{1-\beta_i^2}}{n_i\alpha_i}\cos^{-1}\beta_i \tag{23}$$

So, used as a simple simulation model of a supply chain (with $\alpha$ and $\beta$ chosen to approximate the real decision rules of production managers), these results show explicitly the mechanism by which increasing the system delays ($n_i$) can lead to instability, even in the absence of any logistical disturbances. This occurs if , for any $i$, $\frac{-\sqrt{1-\beta_i^2}}{n_i\alpha_i}\cos^{-1}\beta_i \le h$. In fact, we show later that merely as $\frac{-\sqrt{1-\beta_i^2}}{n_i\alpha_i}\cos^{-1}\beta \to h$ from above, damaging oscillatory production rates ensue.

Consider a supply chain comprising production-inventory systems designed using the (possibly different) approaches presented in section three. In the absence of logistical disturbances (23) tells us that such chains are inherently stable since the individual $h$ – stability margins of figure 4 still apply to the individual systems. The calculation of these quantities proceeds as follows

$$\alpha_i = \frac{K_i^1}{n_i h},\ V_i = \frac{K_i^2}{n_i\alpha_i}, \tag{24}$$

where the constants $K_i^1, K_i^2$ depend on the individual designs in figure 4 and the formula (23). Hence, from (24),

$$V_i = h\frac{K_i^2}{K_i^1} \tag{25}$$

and the system is stable because each design ensures that $K_i^2/K_i^1 >> 1$. However, suppose logistical disturbances (e.g. machine breakdowns, absent workers) increase the delay in the *ith* echelon. Then there would be two different values of $n_i$ in (24), $n_i^{design}$ and $n_i^{actual}$, and so

$$V_i = h\frac{K_i^2}{K_i^1}\frac{n_i^{design}}{n_i^{actual}} \tag{26}$$

So if, for a period of time, $n_i^{actual} / n_i^{design} \geq K_i^2 / K_i^1$ then demand amplification will occur. In fact, we have observed that damaging periods of oscillatory production occur merely if

$$n_i^{actual} \to \frac{K_i^2 n_i^{design}}{K_i^1} \tag{27}$$

from above. The following 4-echelon example illustrates this phenomenon.

*Example*

Consider a 4-echelon supply chain comprising production units with the time delays, from top to bottom, 1.5 days, 2 days, 3 days, 5 days. Echelon 1 might be considered to be the retailer and so then its 'production delay' of 1.5 days would be dominated by the lead time from ordering finished goods to their delivery by echelon 2. Echelon 2 might be a packer/filler, echelon 3 the manufacturer and echelon 4 a raw material supplier. A real supply chain would probably have more raw material suppliers, forming a supply network. Our example follows one path in such a network but can easily be extended to other possible paths. Assume the production managers in each facility have chosen one of our 4 system designs, based on their own perceived requirements. These are (again from top to bottom) designs 1, 4, 3 and 2. Raw material suppliers tend to aim for high throughput utilising near full capacity with few changes in the production rate which might increase downtimes. Yet they typically also aim to hold a large stocks of the finished product, which may become obsolete. So we suppose echelon 4 has chosen design 2, which compromises on its ability to smooth production changes and recover swiftly from inventory discrepancies (see table 2). Echelon 1, a retailer, may possess a greater capacity to hold stocks either on the shelves or in regional depots depending on the product. Since its 'production process' is merely ordering goods, possibly incurring a fixed cost each time it does so,

it may wish to smooth ordering rates and so opt for design 1. In contrast, we might suppose that the packer/filler is keen to be responsive to changes in demand passed down from the retailer (perhaps a supermarket) but simultaneously reluctant to hold expensive amounts of the finished product. So we suppose echelon 2 has chosen design 4, which is most responsive to demand changes and incurs the least inventory depletion. Figure 7 shows the response of the supply chain to a sinusoidal type blip in demand.

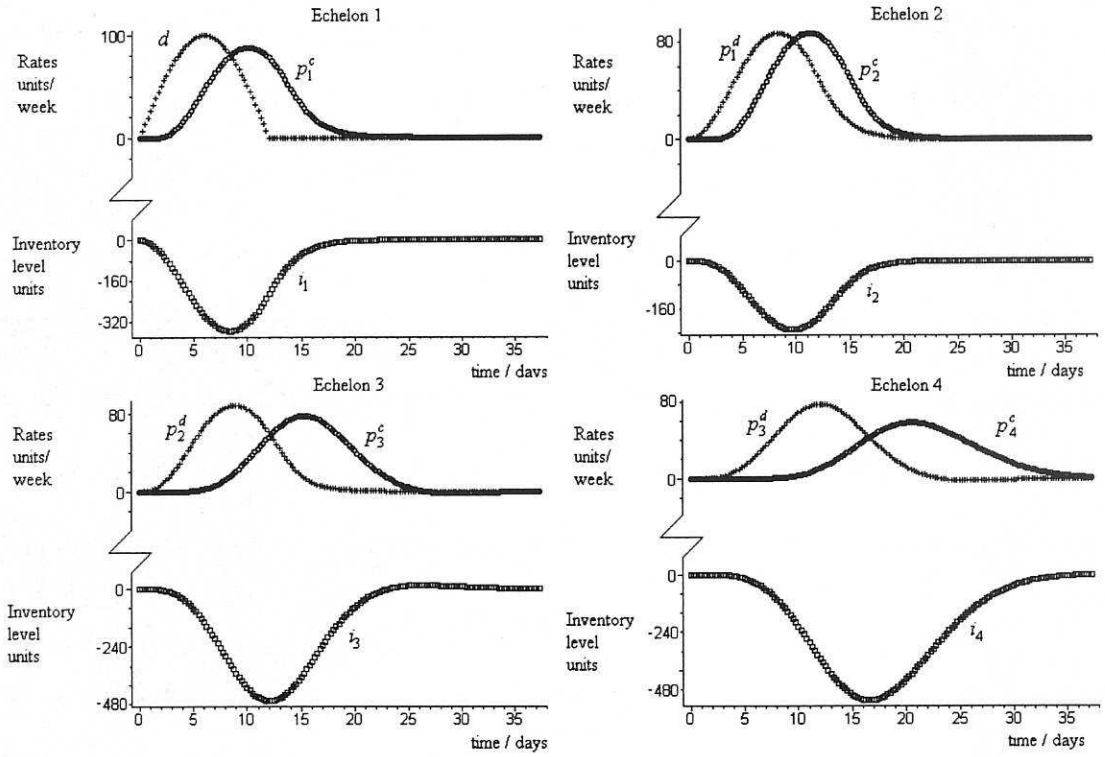


Figure 9. The response of a supply chain to a change in demand

Notice how the cumulative effect of the time delays has stretched the timescale of a transient demand lasting 12 days into a transient production phase lasting almost 30 days in echelon 4. Also, the cumulative smoothing effect of all the designs has amplified the maximum inventory depletion from 360 units in echelon 1 to 550 units in echelon 4, possibly resulting in stockouts, depending on the latter's safety stock

level. Hence policies which are desirable for some supply chain members can have an undesirable impact upon echelons lower down the supply chain.

We now examine the effect of a logistical disturbance in one of the echelons, causing its time delay to increase. The left hand column of figure 8 shows what happens when, given the same demand signal as in figure 7, after 9 days, the time delay in echelon 2 increases to 3.75 days for a further period of 10 days. Since echelon 1 is unaffected by this disturbance, we omit its dynamics. In reality such a logistical disturbance might cause an increase in the time delays of all the higher echelons, since they are supplied by the echelon with the problem. However, in the absence of knowledge of the composition of the time delays for a particular example (e.g. the supply leadtime component in echelon 1's time delay), it is impossible to forecast the magnitude of such increases. Hence we just look at the case when supply to higher levels is unaffected. The same approach can be adopted for the more complicated case given some real data. The point of this example is merely to show the damaging oscillatory effect of logistical disturbances on even a well designed supply chain and how feedback can be used to mitigate such phenomena.

In figure 8 one can see how the disturbance causes a jump discontinuity in the production rate at day 9 (point a), and a further after the extended delay after 12.75 days (point b). Similar discontinuities occur when the delay returns to its original value (points c and d). In lower echelons the action of the respective designs gradually smooths these discontinuities out. However, comparing with the undisturbed case in figure 7, the dynamics now exhibit some oscillation in every echelon, causing expensive variations

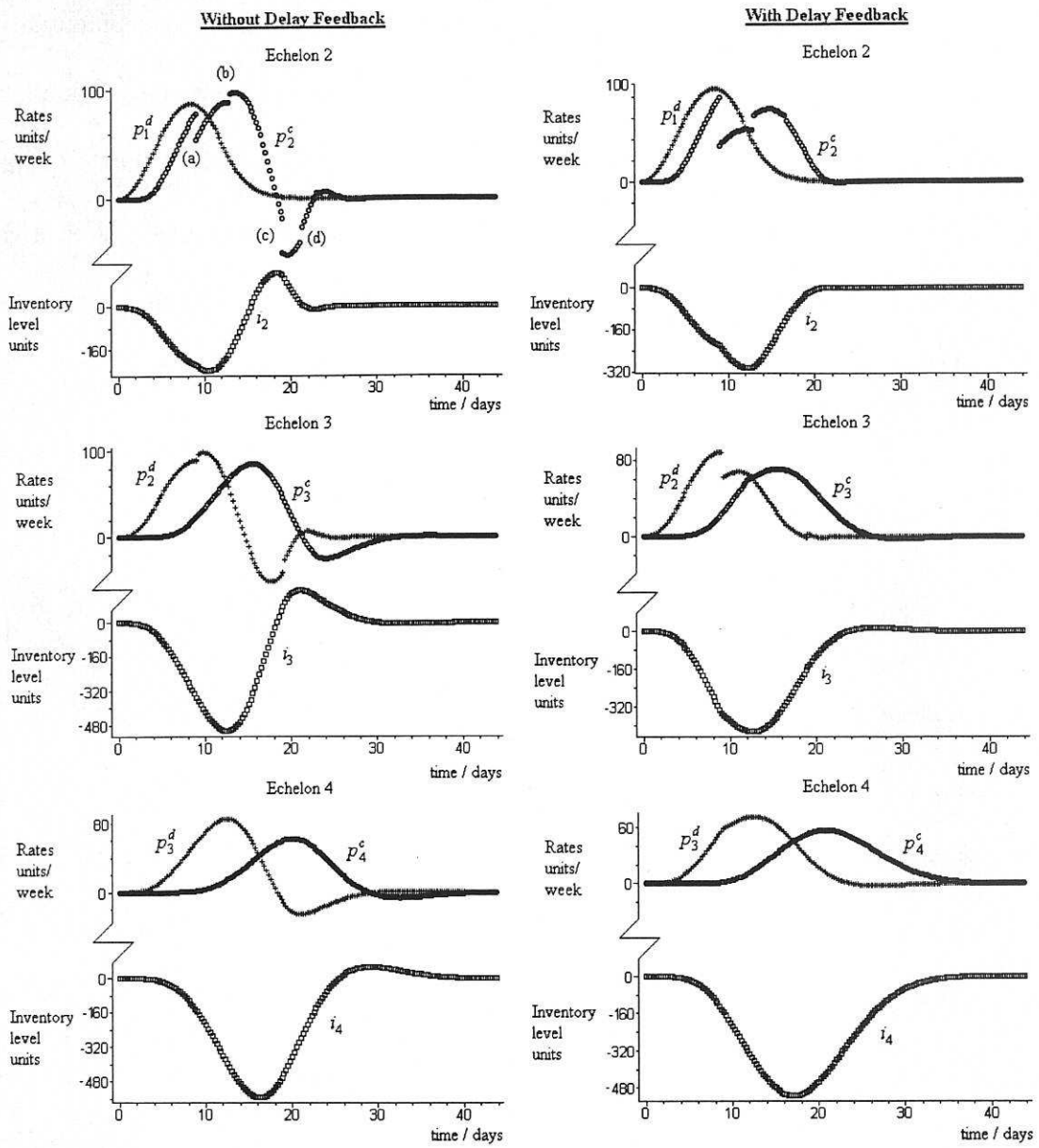


Fig. 10. Response of supply chain to logistical disturbance with and without delay feedback

in production rates over time. Also, the maximum inventory discrepancy in echelons 3 and 4 has increased from –480 to –540 units and –520 to –540 units, respectively, possibly causing stockouts. Furthermore, each echelon now experiences a substantial inventory 'overshoot', periods during which the inventory level exceeds its target. These periods can be very problematic and expensive if enough storage capacity cannot be found and must be rented. It is important to note that the results displayed

above are for well designed systems using rational controller rather than potentially irrational production managers. In the absence of such controllers, the production manager would make decisions which might, quite possibly, lead to much worse oscillations [Sterman, 1989]. Technically, the system is not unstable, rather it is approaching instability. This can be seen by substituting the design rule parameters for echelon 2 (design 4), shown in figure 4, into (24). We have, using (23),

$$h = 0.25 \text{ days}, n_2^{actual} = 15, \mathrm{n}_2^{\text{design}} = 8, K_2^1 = -0.75, K_2^2 = -1.766$$

Instability arises if

$$n_2^{actual} \geq n_2^{design} \, K_2^2 \Big/ K_2^1 = 18.8 \tag{28}$$

The actual value of $n_2^{actual} = 15$ causes oscillatory behaviour due to its proximity to the boundary value calculated in (28). Equation (28) suggests a method for counteracting the undesirable effects of such logistical disturbances through the use of an adaptive controller that monitors the production delay and feeds back this information to adjust the control parameters. The idea is simply to keep the ratio $n_i^{actual} / n_i^{design}$ as close as possible to 1 thus maintaining the particular stability margin shown in table 1. The right hand column of figure 8 shows the results of such a policy, where the design delay exactly matches the actual delay. It should be noted that in reality, at the start of a period of disturbance, the actual delay may not be known and so must be estimated, so a gradual increase in $n_i^{actual}$ might be more realistic. However, changing the parameters of the controller itself can cause discontinuities in the production rate. So, for small disturbances, we recommend relying on any inherent robustness in the original control design rather than its continual tweaking. What 'small' is might only be appreciated by the production manager after using this controller for some time. In

any case, simulation of the effect of adjusting the parameters can be carried out before any adjustment is made.

Figure 8 shows that the feedback eliminates oscillation in all echelons and, in effect, limits the ill effects of the disturbance to echelon 2, whence it originated. The maximum inventory depletion is brought back to the undisturbed level in echelon 4 and, in echelon 2, actually reduced to slightly below this level. The price for this sustained performance is met entirely by echelon 2 through higher maximum inventory depletion. But, even in this echelon, much of the erratic nature of production is suppressed and its peak rate is reduced.

## 5. Conclusions

In this paper we have shown that the approximation of pure production delays by exponential smoothing leads to qualitatively quite different dynamics in a production-inventory system. This may be acceptable when modelling generic supply chain dynamics, but cannot be overlooked when deriving controllers for such systems. By using some straightforward extensions of linear finite-dimensional techniques, we derived four production-inventory system controllers. We then evaluated their performance and, in particular, the balance of the trade-off between smooth production rates, swift inventory depletion recovery and good noise rejection. The robustness properties of these designs were then found using some results on delay-dependent stability conditions.

The second part of the paper examined the dynamics resulting from cascading these production-inventory systems into supply chains. The effect of logistical disturbances that increase the production delay in one echelon was then simulated throughout the

whole chain. It was found that, as instability is approached, costly oscillatory dynamics can occur and, simultaneously, expensive stock depletion is experienced. A heuristic method based on disturbance monitoring and feedback was shown to eliminate such behaviour. Further work is planned on the incorporation of forecasting engines into such systems.

Ansoff, H.I., Slevin, D.P., 1968, Comments on Professor Forrester's industrial dynamics-after the first decade, *Man. Sci.*, **14**, 9, 600.

Bellman, R., Cooke, K.L., 1963, *Differential-difference equations*, (New york: Academic press).

Bensoussan, A., Proth, J.M., 1982, Inventory planning in a deterministic environment continuous time model with concave costs, *Euro. Inst. Adv. Studies in Man.*, working paper, April.

Berry, D., Naim, M.M., Towill, D.R., 1995, Business process re-engineering an electronic products supply chain, *IEE Proc. Sci. Meas. Tech.*, **142**, 5, 395-403.

Bradshaw, A., Porter, B., 1975, Synthesis of control policies for production-inventory tracking system, *Int. J. Systems Sci.*, **6**, 3, 225-232.

Burbidge, J.L., 1984, Automative production control with a simulation capability, Paper presented at IFIP conference, Copenhagen.

Chiasson, J., 1988, A method for computing the interval of delay values for which a differential-delay system is stable, *IEEE Trans. Aut. Con.*, **33**, 12, 1176-1178.

Gorecki, H., Fuksa, S., Grabowski, P., Korytowski, A., 1989, *Analysis and synthesis of time delay systems*, (Chichester: Wiley).

Houlihan, J.B., 1987, International supply chain management, *Int. J. Pys. Dist. Mat. Man.*, **17**, 2, 51-66.

Elsayed, A.E., Boucher, T.O., 1985, *Analysis and control of production systems*, (New Jersey: Prentice-Hall).

Fliess, M., Mounier, H., 1998, Quasi-finite linear delay systems: theory and applications, in *Linear time delay systems*, *Proc. IFAC workshop*, Grenoble, 169-174.

Forrester, J.W., 1961, *Industrial dynamics*, (Cambridge Mass: MIT press).

Forrester, J.W., 1973, *World Dynamics*, (Cambridge Mass: MIT press).

Lieber, Z., 1973, An extension to Modigliani and Hohn's planning horizons results, *Man. Sci.*, **20**, 3, 319-330.

MacDonald, N., 1978, *Time lags in biological models*, (New York: Springer-Verlag).

Mak, K.L., Bradshaw, A. Porter, B., 1976, Stabilizability of production-inventory systems with retarded control policies, *Int. J. Sys. Sci.*, **7**, 3, 277-288.

Marshall, J.E. 1979, *Control of time delay systems*, (Stevenage: Peter Peregrinus).

Marsik, J. 1958, Eine schnelle berechnung des regelungsoptimums nach oldenbourg und sartorius, auch für regelkreise mit transzendenten frequenzgang, *Regelungstechnik*, **6**, 217-219.

Naddor, E., 1966, *Inventory systems*, (Florida: Robert E. Kreiger).

Porter, B., Taylor, F., 1972, Modal control of production-inventory systems, *Int. J. Systems Sci.*, **3**, 3, 325-331.

Ogata, K., 1990, *Modern control engineering*, (Prentice-Hall: New Jersey).

Simon, H.A., 1952, On the application of servomechanism theory in the study of production control, *Econometrica*, **20**, 247-268.

Sterman J.D., 1989, Modelling managerial behaviour: misinterpretations of feedback in a dynamic decision-making experiment, *Man. Sci.*, **355**, 3, 321-339.

Towill, D.R., 1982, Dynamic analysis of an inventory and order based production control system, *Int. J. Prod. Res.*, **20**, 6, 671-687.

Towill, D.R., 1991, Supply chain dynamics, *Int. J. Comp. Int. Manuf.*, **4**, 4, 197-208.

Towill, D.R., 1992, Supply chain dynamics-the change engineering challenge of the mid 1990's, *Proc. Instn. Mech. Engrs.*, **206**, 233-245.

Towill, D.R., Del Vecchio, A., 1994, The application of filter theory to the study of supply chain dynamics, *Prod. Plan. & Control*, **5**, 1, 82-96.

Wikner, J, Towill, D.R., Naim, M.M., 1991, Smoothing supply chain dynamics, *Int. J. Prod. Econ.*, **22**, 231-248.

## APPENDIX

*Derivation of delay-dependent stability criteria for N-echelon supply chain*

Following the work and notation of Chiasson [1988], we examine the unforced system:

$$\frac{d^N}{dt^N}x(t)+\sum_{i=0}^{N-1}\sum_{n=0}^{m}a_{in}\frac{d^i}{dt^i}x(t-nh)=0 \qquad (1)$$

The stability properties of such a system forced by an external function are determined by the roots of the characteristic equation of (1) [Bellman & Cooke, 1963], which, putting $d=e^{-sh}$, we define as

$$a(s,d)=s^N+a_{N-1}(d)s^{N-1}+\cdots+a_0(d),$$

where the $a_i(d), i=0\ldots N-1$, are polynomials. For fixed $h$, (1) is asymptotically stable if $a(s,d)\neq 0$ for $\mathrm{Re}(s)\geq 0, h\geq 0$. Now, define

$$\tilde{a}(s,d)=d^m a(-s,1/d),\quad m=\deg_d\{a(s,d)\}$$

Then,

*Definition 1*: $h^*$, the stability range for $h$.

Let $\{(s_i, d_i) i = 1,\ldots,k\}$ be the common zeros of $\{a(s,d), \tilde{a}(s,d)\}$, for which $\mathrm{Re}(s_i) = 0, s_i \neq 0$ and $|d_i| = 1, d \neq 1$. For each pair $(s_i, d_i)$, let $h_i = \min_{h>0}\{h \in \mathbf{R} | d_i = e^{-hs_i}\}$. Define $h^* = \min\{h_i\} > 0$.

*Theorem 1 [Chiasson, 1988]:*

$$a(s, e^{-sh}) \neq 0 \quad \mathrm{Re}(s) \geq 0 \quad 0 \leq h \leq h^* \tag{2}$$

$$\textit{iff } a(s,1) \neq 0 \quad \mathrm{Re}(s) \geq 0. \tag{3}$$

Further, there exists an $s^* = j\omega^*$ for which

$$a(s^*, e^{-h^* s^*}) = 0$$

thus $h^*$ is the largest value for which (2) holds.

Neutral systems have characteristic polynomials of the form $a(s,d) = a_N(d)s^N + a_{N-1}(d)s^{N-1} + \cdots + a_0(d)$, and are said to be asymptotically stable if $\exists \gamma > 0$ with $a(s, e^{-sh}) \neq 0, \mathrm{Re}(s) > -\gamma$. The reason for this stability margin [Bellman & Cooke, 1963] is that there may exist some unstable solutions of neutral systems for which $a(s, e^{-sh}) \neq 0 \, \mathrm{Re}(s) \geq 0$. So Theorem 1 must augmented by the assumption that $a_N(d) \neq 0 \, |d| \leq 1$.

We now apply this result to the neutral system (i.e. the case where $\beta_i \neq 0$) described by (22). The easier case of $\beta_i = 0$ requires a straightforward application of Theorem 1 and should be obvious in the light of the following. The N-echelon supply chain gives

$$\begin{aligned} a(s,d) &= \{(\alpha_1 + \beta_1 s)d^{n_1} - s\}\{(\alpha_2 + \beta_2 s)d^{n_2} - s\}.....\{(\alpha_N + \beta_N s)d^{n_N} - s\} \\ \tilde{a}(s,d) &= \{(\alpha_1 - \beta_1 s) + d^{n_1} s\}\{(\alpha_2 - \beta_2 s) + d^{n_2} s\}.....\{(\alpha_N - \beta_N s) + d^{n_N} s\} \end{aligned} \tag{4}$$

Now $a_N(d)=(\beta_1 d^{n_1}-1)\ldots(\beta_N d^{n_N}-1)$, which has roots $d_i=\sqrt[n_i]{1/\beta_i}, i=1\ldots N$. Hence the special assumption for neutral systems translates into $|\beta_i|<1, i=1\ldots N$. Now check (3):

$$a(s,1)=\{\alpha_1+(\beta_1-1)s\}\ldots\{\alpha_N+(\beta_N-1)s\}$$

which has roots $s_i=\alpha_i/(1-\beta_i), i=1\ldots N$. Hence, since $\alpha_i<0$, to satisfy (3), we need $\beta_i<1$, a requirement already fulfilled.

The common zeros of (1) and (2) for $|d|=1$ are

$$s_i=\frac{\alpha_i d^{n_i}}{1-\beta_i d^{n_i}},\quad s_i=\frac{\alpha_i}{\beta_i-d^{n_i}},\quad i=1\ldots N$$

respectively. Making them simultaneous gives

$$\beta_i=\frac{1}{2}\left(d^{n_i}+d^{-n_i}\right)$$

Putting $d=e^{j\theta_i}$ gives $\theta_i=\frac{1}{n_i}\cos^{-1}(\beta_i)$. Making the solutions for $d$ equal gives

$$s_i=\pm j\frac{\alpha_i}{\sqrt{1-\beta_i^2}}$$

Now, following Theorem 1,

$$\begin{aligned}h^* &= \min_i \min_{h>0}\left\{h\in\mathbf{R}\mid d_i=e^{-hs_i}\right\}\\ &= \min_i \min_{h>0}\left\{h\in\mathbf{R}\middle| e^{j\theta_i}=\exp\left(\mp jh\frac{\alpha_i}{\sqrt{1-\beta_i^2}}\right)\right\}\\ &= \min_i \min_{h>0}\left\{h\in\mathbf{R}\middle|\theta_i=\mp h\frac{\alpha_i}{\sqrt{1-\beta_i^2}}\right\}\end{aligned}$$

Now, since we are looking for the minimum value of $h$, we use the principle value of $\theta_i$. Hence:

$$h^*=\min_i\left\{-\frac{\sqrt{1-\beta_i^2}}{\alpha_i}\cos^{-1}\beta_i\right\}.$$