Monograph:
Riddalls, C.E. and Bennett, S. (2000) The Stability of Supply Chains. Research Report. ACSE Research Report 779 . Department of Automatic Control and Systems Engineering

Research Report No.779 – September 2000

The stability of supply chains – by C E Riddalls and S Bennett

ABSTRACT:

A continuous time version of the well known beer game model is derived and its stability and robust stability properties are investigated. Novelty originates from the treatment of pure process delays rather than exponential lags and it is shown that this can lead to diametrically different dynamics to the exponential lag case. The stability properties of the system are shown to support and quantify the qualitative empirical results of the beer game. Additional insight into the influence of certain model parameters is attained by their interpretation as the degree of mismatch in a Smith predictor regulator. The transient inability to supply all that is demanded is mimicked and shown to constitute an influential source of demand amplification. The analytical nature of these calculations engenders the capacity to improve supply chain dynamics through the synthesis and calibration of strategic supply chain trade off problems.

THE STABILITY OF SUPPLY CHAINS*

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: A continuous time version of the well known beer game model is derived and its stability and robust stability properties are investigated. Novelty originates from the treatment of pure process delays rather than exponential lags and it is shown that this can lead to diametrically different dynamics to the exponential lag case. The stability properties of the system are shown to support and quantify the qualitative empirical results of the beer game. Additional insight into the influence of certain model parameters is attained by their interpretation as the degree of mismatch in a Smith predictor regulator. The transient inability to supply all that is demanded is mimicked and shown to constitute an influential source of demand amplification. The analytical nature of these calculations engenders the capacity to improve supply chain dynamics through the synthesis and calibration of strategic supply chain trade off problems.

1) Introduction

Since the 1960s a great deal of research has been devoted to the modelling and simulation of supply chains using differential equations (Forrester, 1961). Simplifying these models and omitting their non-linearities, has more recently rendered these models amenable to analytical study (Towill and Del Vecchio, 1994). Consequently the causes of and remedies for such phenomena as demand amplification (Wikner et al., 1991) have been investigated. The merits of this approach are well documented but less so are its deficiencies (Riddalls and Bennett, 2000), chief among which is the absence of any pure time delay component in the production delay. In this paper we explain the differences between time lag systems, which have been widely used instead, and pure time delay systems, and then characterise those real systems which, we feel, are better modelled using pure delays. We then present a comprehensive stability analysis of such systems and demonstrate their substantial qualitative dynamical differences to the standard time lag model. By considering time varying delays this approach reveals an important contributor to the demand amplification effect overlooked by linear time lag (rather than time delay) models. Varying the time delay enables the modelling of the transient inability of suppliers to satisfy all the demand in a given period, a scenario neglected in previous research of this nature. This engenders feedforward (i.e. towards the retailer) effects which have an important influence on the stability of the supply chain as a whole.
The adoption of pure delays creates a continuous version of the general model of managerial behaviour studied in (Sterman, 1989). Our simulations are shown to replicate the qualitative empirical behaviour exhibited by the experiments in this

* This work was supported by EPSRC under the grant number GR/M27432

paper which thus serve as their validation. Extending Sterman's work, we then use the stability properties derived to quantify the potential dynamic improvements possible through the fine adjustment of managers' instinctive production-inventory policies.

The next section gives some background and highlights some of the prominent differences between pure delay and time lag systems, giving our opinion of when each is best used. In section three we introduce the model and simulate some typical responses to standard demand changes. In section four we derive the stability properties of the model and thus relate, in a calibrated way, the results of management decision rules to undesirable performance. By endowing intuitive and anecdotally known effects (e.g. decreasing time delays improves responsiveness) with a quantifiable scale we provide a dynamic dimension to inform strategic decision making in the supply chain. In addition, our analysis throws up some less intuitive results hitherto neglected by similar dynamic models. To conclude, the aim of the paper is to investigate analytically the dynamic consequences of various managerial decision making rules of thumb used to regulate production.

2) Background

The development of Industrial Dynamics was influenced to a great extent by J. Forrester. In his book (Forrester, 1961) he derived a highly detailed nonlinear model of a supply chain using the repeated coupling of first-order differential equation systems. He used the model to simulate 'what if' scenarios but, discouraged by its complexity, carried out no rigorous sensitivity analysis. Such an analysis would have uncovered a high degree of parameter sensitivity in the exponential lags, engendered by the repeated coupling of similar submodels. Forrester was later criticised for the lack of theoretical foundations in these models (Ansoff and Slevin, 1968) and later for similar work modelling world populations (Forrester, 1973). However, simplifying Forrester's models by disregarding their many constraints, Denis Towill and his colleagues facilitated a greater level of analysis whilst still capturing the salient attributes of the system behaviour (Towill 1991, 1992). This approach is used to construct simulation models to evaluate the consequences of re-engineering a fictitious company in (Lewis et al., 1995) and a steel industry supply chain in (Hafeez et al., 1996). The model was also used in studies on the pathology of the demand amplification effect (Wikner et al. 1991, Towill 1991). The analytical tractability of the models was exploited to a fuller extent in (Towill and Del Vecchio, 1994) where a supply chain was regarded as a series of amplifiers. Using classical frequency domain design techniques, each echelon's time lag parameters was tuned to yield a desirable system response. We cite these references to demonstrate the indubitable utility of these models. However, we consider that there are some important issues with these models that have not been investigated and that some clarification and qualification is needed on the extent of the applicability of this approach.

One of the first papers on the design of production-inventory systems (Simon, 1952) highlighted a fact which persists in influencing research today, namely that pure delays are much less amenable to analytical study than their exponential lag approximations. Pure delay systems are fundamentally different since they are infinite dimensional in nature. Their solution can no longer be written in a closed form (in general) and stability analysis becomes the onerous task of locating the position of an infinite number of system poles (Bellman and Cooke, 1963).

Consider two production facilities described by a first order exponential lag and a pure time delay, respectively. The process may describe a combination of activities:

Process delay = raw material delivery leadtime + machine setup time + production time

Suppose the demands placed upon the facility can be described as a function of time, $p^d(t)$, and the resulting production completion rate (the rate at which the finished goods become available to replenish inventory) is $p^c(t)$. The two descriptions of the production process are:

$$\frac{dp^c(t)}{dt} = \frac{1}{T}\left[p^d(t) - p^c(t)\right], \quad p^c(t) = p^d(t - T) \tag{1}$$

and the resulting Laplace domain descriptions are:

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

Simon (1952) justified the use of distributed lags (also called exponential smoothing or exponential lags) by regarding its Laplace transform 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. In contrast Towill has described the lag parameter as a 'production delay of so-many weeks' (Towill, 1991). This is misleading since the actual delay imparted to each entity in the system is different, whereas for pure delays they are all the same. Figure 1 shows the response of the two systems when $p^c(t)$ is a transient step (graph A) and a sinusoidal blip (graph B). One can see there is a great qualitative difference between the response of the time lag and time delay system. Notice also the differing attenuating effect of the exponential lag 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 it is well known that classical design and analysis techniques can be used on systems with small pure delays, so the lag approximation in this case is unwarranted (Marshall, 1979).

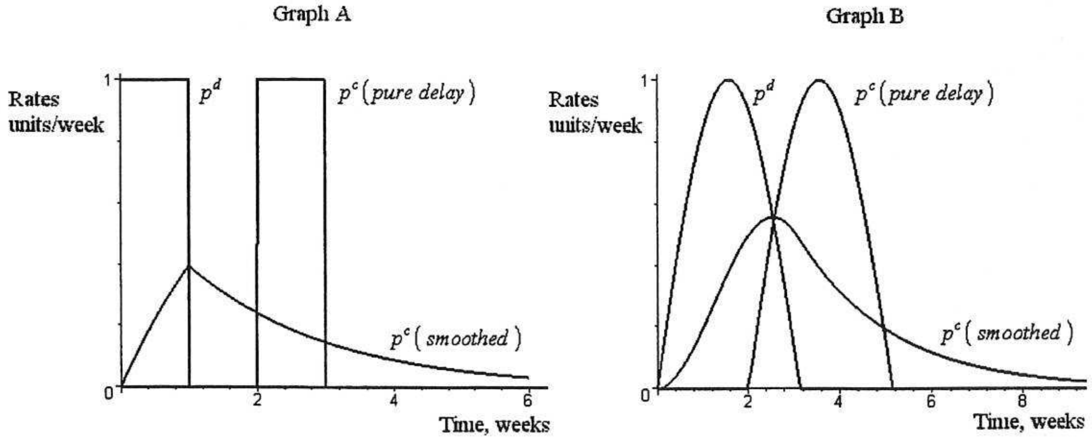

Figure 1. Pure and exponentially smoothed delays

Having established the qualitative differences between lag and delay systems, we now clarify the area of applicability of each. The aggregation of different production lines into a single quantity is often cited as a justification for assuming continuous decision rules in differential equation models of supply chains (Sterman 1987, Forrester 1961). Extending this concept, the distributed lag attributes a different delay to each entity passing through the system. At the machine level of the system this is an attractive argument since after scheduling a job, entities pass through a machine individually, the last having a longer production delay than the first (Incidentally, the inclusion of a pure delay to denote the machine setup time might also be warranted). However, consider graph A over a much shorter timescale spanning a likely period of production. The shape of the completed production curve is unrealistic since it tails off gradually rather than producing at a constant rate then ending abruptly as in real life. Since the system is linear, what is claimed to represent an aggregated quantity should also work for a single production line and first order approximations are demonstrably deficient in this respect. 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. In any case, the timescales used in Towill's models (Hafeez et al., 1996) suggest that he is not regarding each echelon at the detailed machine level. Indeed in models of supply chains rather than individual production-inventory systems the interaction between firms, rather than the detailed description of each companies production process, is more important. In this type of model the production facility's delay should include any leadtime in acquiring raw materials from lower echelons (since in the model the demand is instantaneously transmitted to the lower echelon) and this component tends to be contractually fixed. Hence pure time delays are more realistic in this case and, if the production (machining) leadtime is comparatively small, it may be neglected. Indeed for distribution systems (Sterman, 1989) a pure delay leadtime is unquestionably the most appropriate. Table 1 shows our opinion of the appropriate use of each method.

Type of process delay	lag	lag + pure	pure
Type of process	long production delay	significant machine setup delay and/or significant raw material delivery leadtime	distribution systems long raw material delivery leadtime

Table 1. Suitability of process delays

An important failing of the simplified time lag as used by Towill emanates from the inherent lack of influence each echelon possesses over the next higher. For example, in a typical three echelon model, since all orders placed upon the factory echelon are assumed to be satisfied (after some exponential delay), the internal factory dynamics have no bearing on the system behaviour (see Towill and Del Vecchio, 1994). An important phenomenon - that of the inability to supply all that is demanded over a certain period - is absent from these models. Importantly, research exists that attributes part of the demand amplification effect to this very issue: The rationing game (Lee et al., 1997) occurs during such periods and results in retailers issuing inflated orders in the expectation that they will not be fully met. In the current framework the only way to model inability to supply all that is required is to introduce a nonlinear component into the model. This would hamper analytical progress and perhaps again render the model under the classification of a simulation model. Even when the process delay is entirely made up by production leadtime, the eventuality of increased lags has not been examined. This is an important failing since logistical disturbances to such systems (e.g. absent workers, machine breakdowns) will be manifested in increased production delays. To the Authors' knowledge, no rigorous sensitivity analysis has been carried out on these models to measure the impact of varying production delays during simulation. In contrast, the pure delay systems considered here enable the simulation of inability to supply through their variation over time, whilst preserving the linearity of the system. We also present a robustness analysis which quantifies the effects of these variations on system stability. In the next section, by comparing such properties to those of the lag model, we find diametrically differing results which reveal the inability of the latter models to reproduce such an important and well documented source of demand amplification (Sterman, 1989).

3) Development of the Model

In this section we derive the equations that model a single echelon production-inventory system. By relating the variables to those in (Sterman, 1989), we attain the greater generality of the beer game model enabling us to consider both distribution and production-distribution systems. Although we shall only consider single echelon systems, for clarity of exposition and with the understanding that the behaviour of each contributes to the global dynamics of a supply chain, since the system is linear, the conclusions of this study are pertinent to supply chains which are simply cascaded collections of such models. As we have said, by varying the process time delay we shall mimic the influence of lower echelons on higher echelons without simulating the whole supply chain. Table 2 compares the variables in our model to those found in the beer game.

Production-distribution model		Beer game model	
Inventory	$i(t)$	Inventory	$S(t)$
Demanded production rate	$p^d(t)$	Order rate	$O(t)$
Completed production rate	$p^c(t)$	Acquisition rate	$A(t)$
Demand	$d(t)$	Losses	$L(t)$
Inventory adjustment parameter	α^i	Stock adjustment parameter	α_S
WIP correction parameter	α^{WIP}	Supply line adjustment rate	α_{SL}

Table 2. Comparison of variables in production-distribution and beer game models

Any such model must be based around the fundamental inventory balance equation (3):

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

which is simply a description of the physical flow of entities into and out of the inventory stock. In a distribution system the entities arrive from suppliers and in a production system they roll off the production line and into the inventory store. The demanded production rate, $p^d(t)$, represents the planned or scheduled production rate or, analogously, the rate at which orders are placed upon lower echelons in a distribution system. When there are no logistical disturbances to the system, or equivalently, all orders are met after the delivery leadtime we have the relation.

$$p^c(t) = p^d(t-h) \tag{4}$$

where h is the production leadtime (perhaps including raw material delivery) in a production-distribution system or the delivery leadtime in a distribution system.
We now formulate the decision rule hypothesised in (Sterman, 1989) which controls the order rate through a heuristic that represents local rational decisions by managers. Empirical work in Sterman's paper using a role playing experiment supports such a paradigm. For the general production-distribution system it is postulated that desired production comprises the following components

$$p^d(t) = \underbrace{\hat{L}(t)}_{\text{Expected demand}} + \underbrace{\alpha^i\left[\bar{i}(t) - i(t)\right]}_{\text{Inventory discrepancy term}} + \underbrace{\alpha^{WIP}\left[\hat{h}.\hat{L}(t) - \int_{t-h}^{t} p^d(s)ds\right]}_{\text{Work in progress term}} \tag{5}$$

where $\bar{i}(t)$ is the desired inventory level which, without loss of generality since the system is linear, can be taken to be zero. $\alpha^i \geq 0$ denotes the proportion of the inventory discrepancy (from $\bar{i}$) that is ordered. The work in progress term takes into account those goods that are already in production (or on order) but are not finished (or delivered). This quantity is the integral term, whilst desired WIP is the product of expected production (delivery) delay ($\hat{h}$) and the desired throughput (taken to be equal to the expected demand), $\hat{L}(t)$. A proportion, $\alpha^{WIP} \geq 0$, of the difference between actual and desired WIP is added to p^d. The parameters α^i, α^{WIP} represent the importance attached instinctively by the production-inventory manager to inventory and WIP, respectively. Expected demand is calculated by a trend detector that averages demand over the past T time units. Sterman uses adaptive expectations which are more difficult to model in a continuous time framework. However, our trend detector is widely used in industry to forecast short term demand trends.

$$\hat{L}(t) = \frac{1}{T}\int_{t-T}^{t} d(s)ds \tag{6}$$

Substitute (4) into (3) and differentiate (5) to arrive at the delay differential equation:

$$\frac{dp^d(t)}{dt} = -\alpha^{WIP} p^d(t) - \left(\alpha^i - \alpha^{WIP}\right)p^d(t-h) + \left(\alpha^{WIP}\hat{h} + 1\right)\frac{d\hat{L}}{dt} + \alpha^i d(t) \tag{7}$$

Now, differentiate (6) and substitute into (7) to give

$$\frac{dp^d(t)}{dt} = -\alpha^{WIP} p^d(t) - \left(\alpha^i - \alpha^{WIP}\right)p^d(t-h) + \frac{1}{T}\left(\alpha^{WIP}\hat{h} + 1 + \alpha^i T\right)d(t) - \frac{1}{T}\left(\alpha^{WIP}\hat{h} + 1\right)d(t-T) \tag{8}$$

This equation will form the basis of the stability analysis presented in section 4. Figure 2 shows a Laplace domain representation of the system as modelled in Simulink.

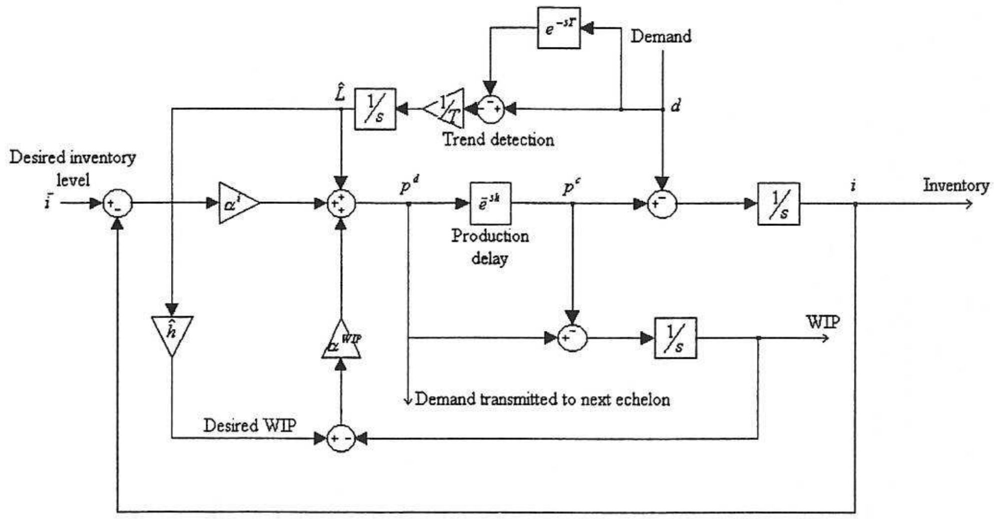

Figure 2. Laplace domain block diagram of production-distribution system

The asymptotic stability of the above system is determined by the parameters α^i, α^{WIP} and the delay length h. Although the other parameters also affect the system response, they do not influence the asymptotic stability (see next section) and so, to avoid confusion, we do not investigate their impact on the model's behaviour here. We shall see the complex interaction between the aforementioned parameters constitutes an important mechanism in itself, and is most relevant to the work in (Sterman, 1989).

Figure 3 shows the response of the system with the following parameters and various combinations of α^i and α^{WIP}.

$$h = 6 \text{ weeks}, \ \hat{h} = 6 \text{ weeks}, \ T = 10 \text{ weeks}, \bar{i} = 200 \text{ units}.$$

As one can see, the demand is a step function (as used by Sterman) of 20% of the initial rate. Since the system is linear, the superposition principle ensures that step functions can be used to approximate any integrable signal and so its choice is entirely general and not calculated to achieve any particular kind of pathological behaviour.

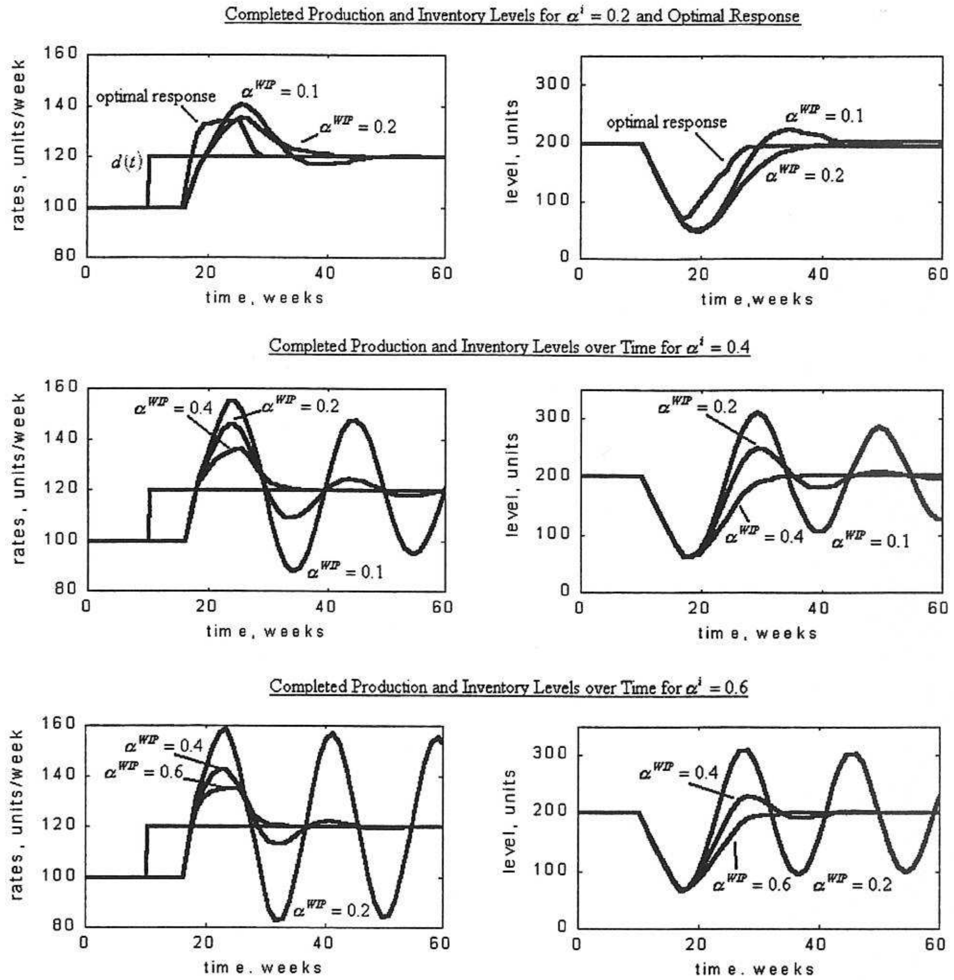

Figure 3. The dynamics of a single echelon system in response to a step increase in demand.

The parameters were chosen to be similar to those found empirically in Sterman's paper. He assumed that $\alpha^{WIP} \le \alpha^{i}$, or that managers always place more emphasis on the inventory level than the supply line (equivalently WIP). This is reasonable since inventory discrepancies are much more immediately apparent to managers than any variance in what is on order. In addition to a persistent underestimation of the importance of WIP, low values of α^{WIP} can be attributed to many factors from an inability to track goods on a production line or in transit, to the effect of information delays and inaccuracies. With this in mind Sterman defines the parameter $\beta = \alpha^{WIP} / \alpha^{i}$ and considers it as the fraction of the supply line taken into account. This is a rather loose interpretation since α^{WIP} itself describes this fraction and if $\alpha^{i} < 1$, but $\beta = 1$, this does not mean the supply line is totally accounted for, merely that it is accounted for in the same proportion as the inventory discrepancy. But, we shall see that β does have an important influence on the system's dynamics.

Returning to figure 3, the top graphs show the optimal response, when $\alpha^{i} = \alpha^{WIP} = 1$, i.e. each discrepancy is entirely accounted for. Notice this simulation exhibits the swiftest response yet the smallest peak and inventory depletion and no oscillatory behaviour. Quite apart from the global supply chain ramifications, oscillation is undesirable because it causes inventory levels to fluctuate and thus creates periods when inventory stocks are surplus to requirements, which increases the storage and handling costs and risk of obsolescence. Similarly it also increases the risk of stockouts when inventory is depleted, thus damaging service levels. On the production side, wildly varying production rates inhibit the most economical scheduling of jobs on machines. Similarly uncertain demand may increase the frequency of reordering raw materials (or the final product in a distribution system) which in turn may increase costs if each order incurs a fixed charge or quantity discounts are not exploited.

Looking at all the responses one concludes that it is not the specific values of α^i and α^{WIP} in isolation that cause oscillatory behaviour, rather their ratio, β. Perhaps this is why it is specifically defined by Sterman in terms of the other parameters. In each case when $\alpha^i = \alpha^{WIP}$ or $\beta = 1$, the response is swift and non-oscillatory. Of course, for smaller values of α^i, the response is slower and the inventory discrepancy thus greater, yet these effect are small compared to the behaviour engendered by even a small difference between α^i and α^{WIP}. In the top two graphs a discrepancy of $\beta = 0.5$ results in slight oscillation, but as this ratio decreases in the lower graphs to 0.25 and 0.33, so too does the amplitude of the oscillation. Sterman concludes, and our simulations demonstrate, that small values of β reflect an incomplete accounting for the 'supply line' leading to transient periods of overordering when demand increases and an eventual surfeit of stock which triggers cuts in production (or ordering) which themselves are similarly disproportionate. This leads to oscillation both in production (or ordering) rates and inventory levels. However given our assertion that α^{WIP} itself, not β measures the level of accounting of the supply line, it is not intuitively clear why this disruptive behaviour occurs irrespective of the particular values of α^i and α^{WIP} and is only sensitive to their ratio. Given our interpretation, when α^{WIP} is small the manager is still prone to over-ordering and yet if $\alpha^i \approx \alpha^{WIP}$ then this does not lead to oscillation. In these circumstances we suggest the reason for this is that the over-ordering still leads to a gradual surfeit of stocks but this is itself under-corrected for in a similar disproportionate manner by α^i, so the dynamics are merely damped symmetrically for both discrepancies. It should be noted that small values of α^i do lead to a slower response and larger inventory depletion and so are undesirable in this respect.

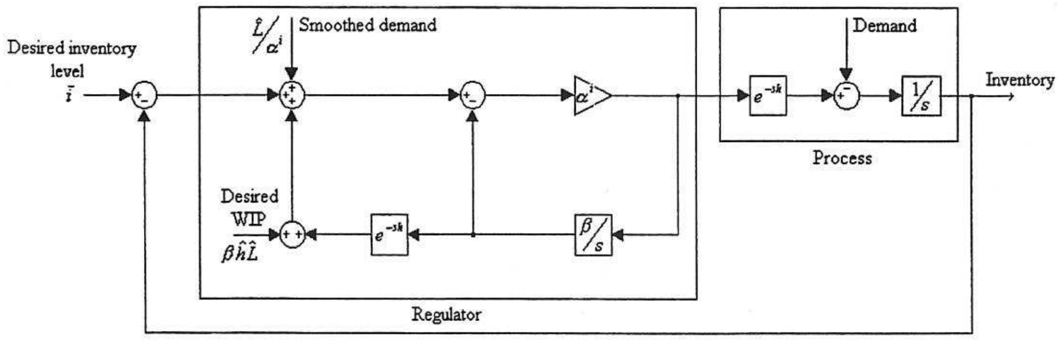

Figure 4. Arrangement of the system in the form of a Smith predictor.

The re-organisation of the block diagram in figure 2 into that shown in figure 4 prompts a new interpretation of the system based on its similarity to the Smith predictor (Smith, 1959). This will afford us a different interpretation for the parameters α^i and β and provide an explanation for the latter's dominant role in determining system stability rooted in well known properties of controllers using the Smith predictor. Smith developed a delay compensation technique in which the regulator (shown in the left hand box in figure 4) uses a mathematical model of the actual process (shown in the right hand box). What this means in our application is that the production manager subconsciously uses his own model of the system to guide decision making. When his cognitive model and the actual process are identical the time delay is removed from the characteristic equation and control can be based on the delay free part of the process (in this case a pure integrator, 1/s). With this interpretation α^i fulfils this role, which accords with our observation that it determines the speed of response. More importantly, β now has a natural interpretation as the measure of the mismatch between the process model and the model use by the regulator (i.e. subconsciously by the production manager). When $\beta = 1$ the two models are equal, but lower values of β signify a degree of mismatch. It is well known (Marshall, 1979) that significant mismatch causes instability via the introduction of a time delay component in the characteristic equation.

In the next section we shall calibrate the exact nature of the stability results demonstrated here and also investigate the complex interaction between the parameter choices, the length of delay and the stability properties. This last area, which was not touched upon by Sterman will prove to be of equal influence to the former.

4) The Stability and Robust stability of Supply Chains

In this section we undertake a comprehensive stability analysis of a single echelon production-distribution system by applying fairly recently derived results on the general stability properties of retarded delay differential equations. Due to the technical nature of this subject we omit any rigorous definition of asymptotic stability and simply refer the interested reader to (Bellman and Cooke, 1963). It will suffice to make do with the concept of stability as that property which brings the system back into equilibrium following a perturbation (e.g. demand change). Conversely instability causes oscillatory behaviour of increasing magnitude over time. In supply chains, technical instability is rare since it is physically impossible for production rates to increase without bound. However, linear dynamical systems approaching instability (in the parameter space) typically exhibit oscillatory behaviour which, in this application, is manifested as demand amplification. So an appreciation of the configurations of parameters (i.e. the particular heuristic used by the manager) that lead to instability should enlighten practitioners to the mechanism of demand amplification and lead to facilitate its attenuation.

The section is split into two parts: Initially we investigate the stability properties of the basic single echelon system with a fixed production/distribution delay. In particular we shall aim to quantify the effect, demonstrated above and in (Sterman, 1989) of the parameter β on stability. We then reveal the influence of the other parameter, α^i, and attempt to provide intuitive explanations for the analytical results derived. Unlike Sterman we then investigate the influence of the magnitude of delays on stability and provide a precise quantification of the well known link between long delays and demand amplification (Towill, 1991). We also demonstrate the diametric qualitative differences between the stability properties of exponential lag systems and the pure delay systems considered here.

Since single echelon systems rarely exist in isolation we extend this study in the second part of the section to examine their robust stability properties. As mentioned in section two, the influences in these models have hitherto been limited to the direction away from the retailer and are solely manifested in the transmitted demand signal. However, in Sterman's paper an important source of instability is cited as the inability to supply all that is demanded. This can be modelled as a transient increase in the time delays involved, a fact which motivates the investigation of the robust stability characteristics of the system. This study reveals a significant source of demand amplification previously neglected in dynamic modelling research of this type. We show that varying time delays reduces the robustness of each echelon and its ability to deal satisfactorily with transient demand blips. Lastly we investigate the consequence of a transient change in the heuristic used by the manager, resorted to in adverse conditions and manifested as a time varying α^i.

Basic Stability Properties

The asymptotic stability of the standard production-inventory system is determined by the characteristic equation of the unforced part of (8):

$$\frac{dp^d(t)}{dt} = -\alpha^{WIP} p^d(t) - \left(\alpha^i - \alpha^{WIP}\right) p^d(t-h) \tag{9}$$

Since we have determined that the parameter β is of such importance, we rewrite (9) thus

$$\frac{dp^d(t)}{dt} = -\alpha^i \left\{\beta p^d(t) + (1-\beta) p^d(t-h)\right\} \tag{10}$$

For delay differential equations there are two types of stability: delay dependent stability and stability independent of delay (IoD). Obviously the latter is more restrictive since it must encompass all those parameter combinations that result in systems that are stable for all delay values. Simple conditions for systems of the form (10) have been known for some time (Bellman and Cooke, 1963) and result in the following:

The system (10) is asymptotically stable IoD if and only if $\beta \geq \frac{1}{2}$.

If, on the other hand, $\beta < \frac{1}{2}$, then the delay dependent stability criterion is:

The maximum delay for which (10) is asymptotically stable is given by

$$h^* = \frac{\arccos\left(\frac{\beta}{\beta - 1}\right)}{\alpha^i \sqrt{1 - 2\beta^2}} \tag{11}$$

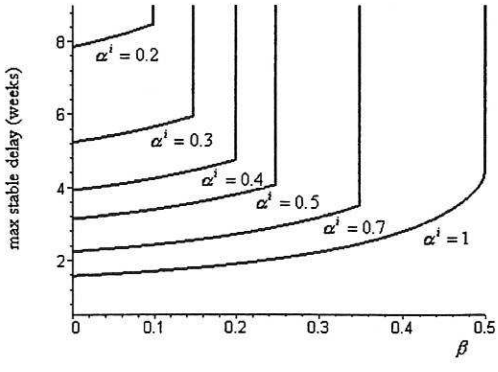

Figure 5. The stability regions for (8) and various parameter combinations.

The stability regions for various α^i and varying β are shown in figure 5 as the area to the right and below each curve. The curves become straight lines when the system is stable IoD. One can see that for smaller values of α^i the system is stable for a much larger range of delays. This is reasonable since smaller values will tend to make the system less responsive to inventory fluctuations. Although the response may be less desirable due to an increased risk of stockouts, damaging oscillations are reduced. Given a particular value of β, this trade off can now be quantified using figure 5.
It is apparent that β plays the most important role in determining stability, especially stability IoD. We can conclude that for robust stable systems, managers should attach at least half as much importance to the supply line (or WIP) as the inventory level. For good dynamic behaviour (i.e. swift response, no overshoot, small inventory discrepancy) we have seen that systems with $\beta = 1$ perform best and are 'most' stable (i.e. furthest away, in the parameter space, from unstable systems). However, small values of α^{WIP} may be beyond the control, in the short term, of managers due to inaccurate data, an inability to track WIP and information delays. For example, if $\alpha^{WIP} < 0.2$, say, then large values of α^i may be damaging to the stability properties of the system since they lead to small values of β which can lead to instability ($\alpha^i > 0.4$ in this case). In conclusion, it is most important to make inventory and WIP adjustments in similar proportions, otherwise one will overcorrect for the other, leading to oscillations.

Comparison with exponential lag models

In (John et al., 1994) the authors examine a similar model which uses exponential lags instead of pure time delays. We have set forth in detail our reasons for preferring the latter for some applications and now reveal the large qualitative differences in the stability properties of each model. The denominator of the equivalent transfer function for $p^d(t)/d(t)$ can be derived from equation (4) in (John et al., 1994) to be

$$(1+T_p s)(1+T_a s)\left(1+\left(1+\frac{T_p}{T_w}\right)T_i s + T_p T_i s^2\right),$$

which, omitting the 2 left hand brackets, since they do not affect the stability, and substituting our parameters, gives

$$\alpha^i + (1+\alpha^i \beta h)s + hs^2 \tag{12}$$

Now, by freezing α^i and β at plausible values, we compare the stability properties of this system with those derived above for the same system with pure delays. For $\alpha^i = 0.5, \beta = 0.2$ plotting the zeros of (12) for various h demonstrates that the exponential lag system is stable for all positive values of the delay parameter and, more importantly becomes more stable (i.e. the real part of each zero becomes more negative) as h increases. This is in stark contrast to our system which, from (11) becomes unstable for $h = 3.88$ weeks. The stability properties of the exponential lag system fails to confirm previous research which shows that long delays create costly oscillatory behaviour through uncertainty (Houlihan, 1987). Therefore, our simulation model has confirmed a well known phenomenon which hitherto has been neglected and indeed cannot be replicated by the exponential lag approach.

Robust stability

We have seen so far how the dynamic response of our system is related to the feedback parameters α^i and β and the magnitude of the delay, h. However, Sterman's empirical work has shown the damaging consequences of an inability to supply all that is demanded over a given period. This leads to over-ordering and a consequent surfeit of stocks when the goods do eventually arrive. To our knowledge, no mechanism to study this important phenomena has thus far been incorporated in the Forrester-Towill model. The substitution of pure delays for exponential lags permits the modelling of these effects by varying the delay magnitude during the simulation period. In production-distribution systems a lengthening of the time delay may also result from an internal logistical disturbance to the production process, like machine breakdowns, absent workers etc. These observations prompt the study of the robust stability characteristics of the system, i.e. to what extent do delay variations push an otherwise stable system towards instability?

The calculation of the equivalent to h^* (11) is more difficult due to the non-stationary nature of the system and an ignorance of the exact nature of the variation. For these reasons the following conditions are only necessary not sufficient for stability and so may not give an optimal bound as in the constant delay case (11). However, they serve as a guide to the tendency of such variations to degrade the system's stability properties, supported, in any case, by examples to illustrate this effect.

(Niculescu et al., 1997) The system (10) with h *replaced by* $h(t)$ *and* $H(h) = \{h(t) \in C^0, 0 \le h(t) \le h, \forall t \in \Re^+\}$, *is stable IoD for* $h(t) \in H(h)$ *if* $\beta > \frac{1}{2}$. *On the other hand, if* $\beta < \frac{1}{2}$ *then the stability is guaranteed if* $h \in H(h^*)$, *where*

$$h^* = \frac{1}{\alpha^i (1-\beta)} \tag{13}$$

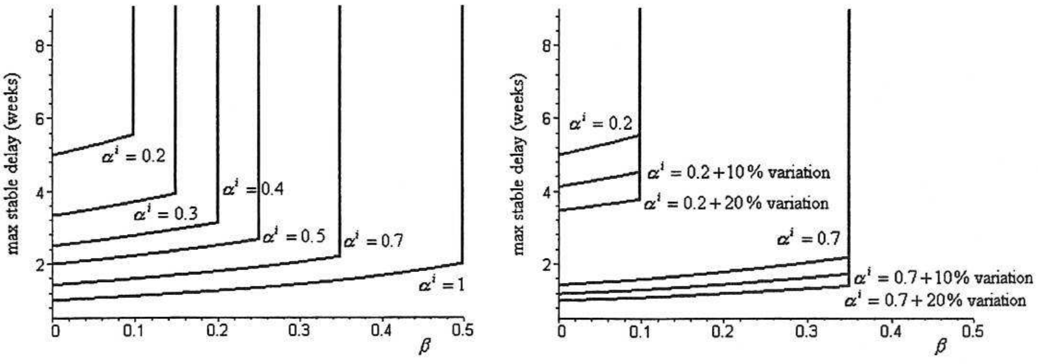

Figure 6. Stability regions for varying delay and α^i

The guaranteed stability region is shown in the left hand plot of figure 6 again as the region to the right and below each curve. Each stability region is considerably reduced compared to the equivalent in figure 5. Consider a system with the following parameters:

$$\alpha^i = 0.7,\ \beta = 0.3,\ h = 3\,\text{weeks},\ \hat{h} = 3\,\text{weeks},\ T = 10\,\text{weeks}, \bar{i} = 200\,\text{units}.$$

Figure 7 shows the response of the system to a triangular type demand blip for both a constant delay and a delay which varies as a sinusoidal blip between weeks 10 and 30 from a magnitude of 3 weeks to a maximum of 5, and then back to 3 weeks.

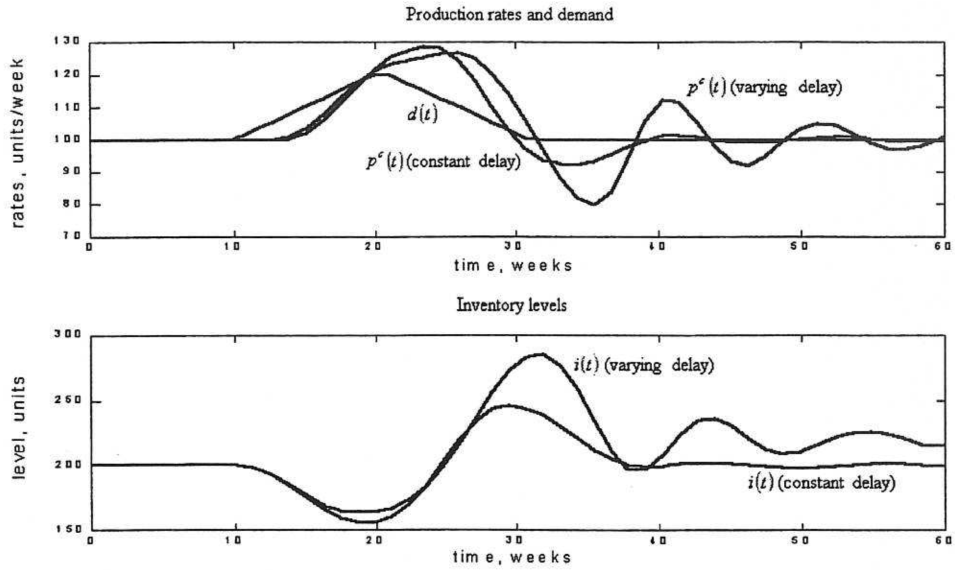

Figure 7. Response of the system with constant and varying delay.

One can see that such a delay variation has caused oscillation in the inventory level and production rate for 30 weeks after it has ceased. These effects are important since they may be caused by factors beyond the immediate control of that echelon, such as a transient lengthening of the raw material delivery leadtime. By quantifying the propensity for oscillations to arise due to transient lengthening of the process time delays, this analysis may be used to inform strategic supply chain decisions. For example, if the supply of raw materials is unreliable it may prove more cost effective to keep a larger safety stock and so ensure production leadimes remain unaffected.

Similarly, the damaging effects (on the whole supply chain) of machine breakdowns can now be fully calculated.

Sterman found that replacing the humans decisions with the model postulated (and used here) increased their performance by around 5%. This he attributed to the consistency of the decision making rules. So there is some evidence to suggest that at any given time the heuristic adopted may vary about a mean parameter set. It is also plausible to assume that managers are particularly liable to such inconsistencies during periods of instability and large inventory discrepancies. To investigate this we derive the stability properties of our system given variations in α^i . Variations in this parameter are considered more likely than in α^{WIP} since inventory oscillations are much more salient and directly burden the echelon with addition costs, whereas a longstanding neglect of WIP or goods on order is unlikely to be rectified during adversity. Suppose $\overline{\alpha}^i = \sup \alpha^i(t)$ and $\gamma = \overline{\alpha}^i / \alpha^i$, is the maximum variation about the notional value α^i. Then

(Niculescu et al., 1997) The system (10) with h *replaced by* $h(t)$ *is stable IoD for* $h(t) \in H(h)$ *if* $\overline{\beta} = \alpha^{WIP} / \overline{\alpha}^i > 1/2$. *On the other hand, if* $\overline{\beta} < 1/2$ *then the stability is guaranteed if* $h \in H(h^*)$, *where*

$$h^* = \frac{1}{\gamma \alpha^i (\gamma - \beta)} \tag{14}$$

The right hand plot in figure 6 shows the reduction in the guaranteed stability regions of the system resulting from such a variation and figure 8 illustrates these dynamically for the varying delay example of figure 7. In an attempt to damp down the large oscillations in inventory levels shown in figure 7 (and, for comparison, figure 8), we suppose the inventory manager accelerates his response by 20% (to $\alpha^i = 0.84$) at week 30.

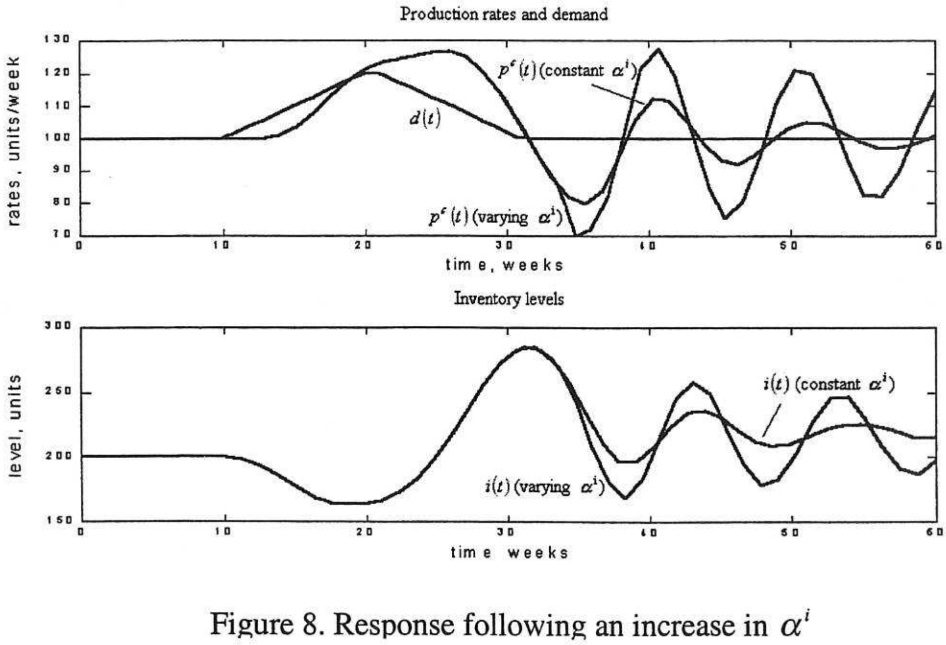

Figure 8. Response following an increase in α^i

One can see that instead of smoothing the response, this increases oscillation significantly since α^{WIP} is not augmented by a similar proportion and the over-ordering is itself over-compensated by a greater margin after the increase in α^{i}. Thus we have shown how plausible well-intentioned action to smooth supply chain dynamics can lead to greater oscillations. We conclude that consistency of decision making is a desirable property.

In this section we have elucidated and quantified the mechanisms by which management decisions can lead to oscillation in the supply chain. The ratio of the two correction parameters, β, emerged as most influential since if it is greater than 0.5 the system remains stable for any delay magnitude. Although Sterman's empirical work has revealed that β is often a lot less than 1 due to a cognitive neglect of the supply line, in the real world this may also be attributable to a range of factors from an inability to track WIP to inaccurate or untimely data. By relating β to the maximum delay magnitude which guarantees stability we have enabled the calculation of the following trade off: Whether to invest in technology to counter the above (e.g. tracking technology) or purchase new machinery or renegotiate shorter raw material leadtimes to decrease h. These three policies may be viewed as increasing the agility of the system. Alternatively we have shown that reducing α^{i} can attenuate oscillatory behaviour at the expense of a slower response and concomitant increased risk of stockouts. Our analysis provides the manager with the capability to decide whether to invest in increased safety stocks and allow a reduction in responsiveness rather than invest in measures to increase agility.

We have also quantified how lower echelons can influence performance through a transient inability to supply all that is required on time. These results furnish the production manager with the information required to gauge whether greater stocks of raw materials should be held. We demonstrated how a transient inability to supply in one echelon can engender oscillatory production patterns in the echelon above and, since this is passed on as transmitted demand, in oscillations further down the supply chain. So the echelon which suffers the temporary stockout is also subject to demand amplification to which it has contributed. By relating this phenomenon to the parametric characteristics of the model we have revealed the conditions under which such a 'vicious circle' may flourish: It is important that systems be stable for a range of delays about the nominal value. Given their grounding in empirical research, such simulation and analytical results could be used to promote supply chain wide strategies for the attenuation of demand amplification.

5) Conclusions

In this paper we have examined the stability properties of a continuous time version of a general model of managerial behaviour as validated in (Sterman, 1989). We showed how the treatment of pure delays as opposed to exponential lags gives rise to qualitatively very different stability properties. In particular we demonstrated the importance of robust stability, i.e. stability for a range of delays, and how stockouts in lower echelons can create a vicious circle of unstable influences in the supply chain. The analytical nature of these calculations enables the possible calibration of strategic trade off decisions for the improvement of supply chain dynamics. The application of the theory developed here to characteristic trade off problems will be presented in a future paper.

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).
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.
Chen, F., 1999, Decentralized supply chains subject to information delays, *Man. Sci.*, **45**, 8, 1076-1090.
Forrester, J.W., 1961, *Industrial dynamics*, (Cambridge Mass: MIT press).
Forrester, J.W., 1973, *World Dynamics*, (Cambridge Mass: MIT press).
Hafeez, K., Griffiths, M., Griffiths, J., Naim, M.M., 1996, Systems design of two-echelon steel industry supply chain, *Int. J. Prod. Econ.*, **45**, 121-130.
Houlihan, J.B., 1987, International supply chain management, *Int. J. Pys. Dist. Mat. Man.*, **17**, 2, 51-66.
John, S., Naim, M.M., Towill, D.R., 1994, Dynamic analysis of WIP compensated decision support system, *Int. J. Man. Sys. Design*, **1**, 4, 283-297.
Lee, H.L., Padmanabhan, V., Whang, S., 1997, Information distortion in a supply chain: the bullwhip effect, Management Science, **43**, 4, 546-558.
Lewis, J.C., Naim, M.M., Towill, D.R., 1995, Re-engineering the supply chain interface-an integrated approach, *2nd Int. Conf. on Logistics*, Nottingham, 329-335.
MacDonald, N., 1978, *Time lags in biological models*, (New York: Springer-Verlag).
Marshall, J.E. 1979, *Control of time delay systems*, (Stevenage: Peter Peregrinus).
Niculescu, S-I., Verriest, E.I., Dugard, L., Dion, J-M., 1997, Stability and robust stability of time-delay system: a guided tour, in *Stability and control of time delay systems (Dugard, L., Verriest, E.I., Eds.)*, LNCIS, (London: Springer-verlag), **228**, 1-71.
Riddalls, C.E., Bennett, S., 2000, Modelling the dynamics of supply chains, *Int. J. Sys. Sci.*, **31**, 8, 969-976.
Simon, H.A., 1952, On the application of servomechanism theory in the study of production control, *Econometrica*, **20**, 247-268.
Sterman, J.D., 1987, Testing behavioural simulation models by direct experiment, *Man. Sci.*, **33**, 12, 1572-1592.
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.