Capacity restrictions and supply chain performance: Modelling and analysing load-dependent lead times

Abstract

Several studies have reported that capacitated supply chains may benefit from an improved dynamic performance as compared to unconstrained ones. This occurs as a consequence of the capacity limit acting as a production smoothing filter. In this research, the relationship between capacity restrictions and the operational performance of supply chains is investigated from a novel perspective, i.e. we assume that the influence of capacity constraints on lead times depends on the supply chain's responsiveness. Under these circumstances, the experiments show that there is a negative effect of capacity constraints on supply chain performance, both in terms of process efficiency (internal) and customer satisfaction (external). Nonetheless, the magnitude of this impact greatly depends on the responsiveness of the firm, market conditions and adopted policies for inventory management. More specifically, the combination of tight capacity restrictions and low responsiveness significantly contributes to decrease supply chain performance, which may be very damaging for the dynamic behavior of the system. In this sense, efficient capacity planning proves to be essential to prevent the supply chain from entering into a vicious circle.

Keywords

Bullwhip Effect; capacity constraints; CT-TP curve; lead time; responsiveness; supply chain dynamics.

1. Introduction.

1.1. Context.

Supply chains are severely impacted by the dynamic phenomenon known as the Bullwhip Effect (Lee et al., 1997), which amplifies the variability of orders as they pass through the various echelons of the system. The first records of this phenomenon by Procter & Gamble date back to the 1910s (Schisgall, 1981). Some decades later, Forrester (1958) initiated its theoretical analysis via system dynamics simulation in the MIT. However, it was not until the 1990s when the Bullwhip Effect became a major stream of research within the management literature, which is possibly a consequence of the new business environment drawn by globalisation. In this increasingly competitive scene, supply chain management has become a key success factor for firms (Buchmeister et al., 2014), while the Bullwhip Effect, which creates a climate of instability in production and distribution systems, tends to decrease the firms’ operational and financial performance (Disney and Lambrecht, 2008).

A direct consequence of the Bullwhip Effect is the generation of highly variable production schedules (Disney and Towill, 2003). Consequently, companies need to invest in high capacity to meet peaks in demand, while this capacity will be later underutilised when demand decreases. In this sense, the Bullwhip Effect closely relates to a traditional problem in the operations management field, namely, the capacity choice (Cachon and Lariviere, 1999). Under these circumstances, the relationship between capacity constraints and the dynamic performance of supply chains has been explored by several authors. Interestingly, some works have observed a positive impact of capacity restrictions on supply chain performance that emerges from these acting as a production smoothing mechanism (Evans and Naim, 1994; Cannella et al., 2008; Chen and Lee, 2012; Ponte et al., 2017). Nonetheless, this operational improvement generally occurs at the expense of decreasing the inventory performance of the system, which in turn results in a decreased customer service level (Evans and Naim, 1994; Nepal et al., 2012; Hussain et al., 2016; Ponte et al., 2017).

1.2. Problem statement.

Despite the evident interest on capacitated supply chains, none of these prior studies have introduced into the analysis a common cause-effect relationship in real-world supply chains: the impact of capacity constraints on the lead time (Upasani and Uzsoy, 2008; Orcun et al. 2009, Fransoo and Lee, 2013; Kacar et al., 2016), as high orders usually increase the time that manufacturers, with constrained capacities, need to replenish these orders (Sterman, 2006; Boute et al., 2007). In other words, decreasing capacity tends to increase lead times throughout the supply chain. Note that this situation may lead to a 'vicious circle': The Bullwhip Effect increases the lead times in the supply chain, which in turn causes higher Bullwhip Effect due to the need of protecting against demand uncertainty for longer periods of time. This circle has been previously identified (Disney and Lambrecht, 2008), but has been barely explored in the literature (Childerhouse et al., 2008).

To the best of the author's knowledge, only three studies explore supply chain performance explicitly assuming load-dependent lead times, namely, Helo et al. (2000), Boute et al. (2009) and Framinan (2017). Helo et al. (2000) employ a lead time factor defined as the ratio of the backlog to the capacity. Boute et al. (2009) use a time queueing model to estimate the lead time distribution. Framinan (2017) models a case in which capacity is directly linked to current/past orders and/or demand (among others). All three studies suggest that capacity imitations may increase operational costs, including those related to the Bullwhip Effect, in countertendency with other studies --such as the aforementioned Evans and Naim, (1994); Cannella et al., (2008); Chen and Lee, (2012); Ponte et al., (2017) -- assuming the rejection of orders in excess of a capacity threshold. These works have provided significant insights regarding modelling and analysis of capacitated supply chains with load-dependent lead time. However, all three studies have been developed under a number of rather restrictive assumptions and thus, the impact of capacity has been studied only for a limited number of market and decision-making scenarios. As a summary, we can conclude that the impact of load-dependent lead times and capacity constrained supply chains has been understudied in the literature of supply chain dynamics (see Section 2).

1.3. Objective.

Motivated by the above-mentioned considerations, we argue that a possible avenue to improve the understanding of the effect of capacity constraints in supply chain performance can be given by the two following objectives:

(1) To explicitly model and analyze the effect of capacity and load-dependent lead time on the basis of empirically-driven assumptions/observations

(2) To explore the impact of capacity restrictions under a variety of scenarios, including supply chain responsiveness, variability of market demand, and replenishment decisions.

To fulfill objective (1) we model the capacity using a cycle time-throughput (CT-TP) curve, a load-dependent lead time model commonly adopted in industry for the estimation of cycle times. More specifically, the CT-TP curve empirically quantifies the relationship between the average cycle time and the throughput rate (Ankenman et al., 2011). By doing so, we provide novel insights regarding the relation between capacity constraints and the operational performance of supply chains by investigating how load-dependent lead times influence the behavior of capacitated supply chains. To fulfill objective (2) we explore the capacitated supply chain for different levels of supply chain responsiveness, variability of market demand and replenishment decisions. More specifically, supply chain responsiveness is considered as the ability of the system in delivering the same product within a shorter lead time. The influence of market demand and replenishment decisions are modelled by considering different levels of turbulence in a customer demand (*i.e.*, coefficient of variation), order policies (*i.e.*, classical order-up-to (OUT) and smoothing OUT) and customer service level (*i.e.*, safety stock factors). We adopt a performance measurement system aimed at capturing both operational costs (*i.e.*, demand amplification and inventory instability) and customer satisfaction (*i.e.*, percentage of delivering products to customer)

Our methodological approach is based on modelling and simulation techniques and supported by inferential statistics (Kleijnen et al. 2008, Evers and Wang 2012.). The capacitated supply chain is modelled and implemented using difference equation (Riddals et al., 2000). The experiments have been carried out according to a full factorial design and their results have been examined by analyzing the main effect of the capacity and its interaction with the other four analyzed factors (*i.e.*, responsiveness, coefficient of variation of customer demand, proportional controller of the OUT and safety stock factor). The results suggest that, in a capacitated supply chain with load-dependent lead time, capacity constraints significantly impact supply chain performance, and that this impact depends on the responsiveness of the supply chain (lead time increase), market conditions (demand variability), and on the replenishment decisions (safety stock factor and proportional controller). In view of the results obtained, we derive three main implications for researchers and practitioners regarding (1) the assumption on capacity in supply chains, (2) the investments on capacity and higher responsiveness of production-distribution system, and (3) the use of some inventory management decisions for limiting the potential negative effect of capacity restrictions.

The rest of this article is structured as follows. Section 2 reviews the most relevant literature on the subject. In Section 3 we detail the capacitated supply chain model, with an especial emphasis on the lead time function, and define the key performance metrics. Section 4 describes the experiments and presents the main results and findings. In Section 5 we discuss the managerial implications reflected from our results. Finally, Section 6 concludes and paves avenues for future research.

2. Literature review: the capacity-constrained supply chain.

Many studies in the supply chain field assume unconstrained production, distribution, and storage capacities, which can be interpreted more as a necessity than as an attempt to model real-world systems (Shukla and Naim, 2017). Although "one of the relevant features of the global enterprise business network is the constrained capacity of production plants and distribution centres" (Ciancimino and Cannella, 2009), some common techniques in this area —such as control engineering or stochastic analysis— present serious difficulties when dealing with nonlinearities (Grubbström and Wang, 2000; Riddalls and Bennett, 2002). For this reason, the works investigating the implications of capacity limits on the supply chain response are relatively scarce and most of them have been carried out using simulation techniques, as highlighted by Cannella et al. (2008) and Ponte et al. (2017). This section is devoted to review these studies, whose the most relevant information is included in Table 1. The first column shows the reference to the article. Then, it follows the methodological approach, the variable over which the constraint was placed, and the assumptions on lead time modelling. Finally, we summarize their main conclusions regarding the impact of capacity restrictions on supply chain performance.

Table 1 shows that the dynamic analysis of the capacity-constrained supply chain was initiated by Tang and Naim (1994). They compared eight different three-echelon supply chains that only differ in the capacity assumptions, and discovered that the unconstrained system did not produce the best response. The operational

Table 1. Relevant research on the operational impact of capacity constraints on supply chains.

[Ref] Authors (Year)	*Methodological approach*	*Capacity limit*	*Lead time assumption*	*Key findings*
[i] Evans and Naim (1994)	Simulation (differential equations modelling)	Order rate	Independent	• Capacity limits tend to decrease inventory service levels, but they generally lead to an improved dynamic performance. • Overall, the unconstrained system does not always produce the best response.
[ii] de Souza et al. (2000)	Simulation (system dynamics)	Production	Independent	• Capacity constraints have a major impact on supply chain dynamics and costs. • The supply chain response is seriously damaged by capacity shortages; hence capacity planning becomes essential.
[iii] Helo (2000)	Simulation (system dynamics)	Production	Load-dependent	• Reduced capacity damages the agility of the supply chain. • Capacity utilization can be used as a substitute for inventory.
[iv] Vlachos and Tagaras (2001)	Mathematical analysis & simulation (differential equations modelling)	Order rate	Independent	• Capacity limits have a negative impact on system performance especially when lead times are long.
[v] Wilson (2007)	Simulation (system dynamics)	Transportation (short-term)	Independent	• Short-term capacity loss due to transportation disruption results in a reduced fill rate, but it may generate a dynamic improvement in the system.

[vi] Cannella et al. (2008)	Simulation (differential equations modelling)	Order rate	Independent	• The Bullwhip Effect is mitigated when capacity is considered in the supply chain. • An increment of production capacity does not necessarily imply an improvement in customer service.
[vii] Boute et al. (2009)	Mathematical analysis	Production	Load-dependent	• Inflexible capacity results in stochastic lead times, thereby increasing the inventory requirements and supply chain costs.
[viii] Ciancimino and Cannella (2009)	Simulation (differential equations modelling)	Production	Independent	• The "rogue" dampening of the Bullwhip Effect provoked by capacity constraints increases supply chain risk, as it may lead to satisfy at a higher cost a "false" demand.
[ix] Juntunen and Juga (2009)	Simulation (discrete event)	Transportation	Independent	• An increase in the transportation capacity does not necessarily translate into an improved customer service.
[x] Hamdouch (2011)	Network equilibrium	Production and shortage	Independent	• Capacity restrictions do not only affect supply chain dynamics but also market response. By impaction on the impact on the price of the product, demand forecasting becomes more complex, which will add to the generation of the Bullwhip Effect.
[xi] Nepal et al. (2012)	Simulation (differential equations modelling)	Production	Independent	• While capacity limits do not significantly impact order variability, they increase inventory variability.
[xii] Chen and Lee (2012)	Mathematical analysis	Order rate	Independent	• Imposing a finite capacity to supply chains smooths the order variability.
[xiii] Spiegler and Naim (2014)	Simulation (system dynamics)	Transportation	Independent	• Capacity limitations negatively impact both inventory and backlog costs, although there is a positive impact on the

				'backlash' effect (the reflection of the Bullwhip Effect on transportation).
[xiv] Hussain et al. (2016)	Simulation (system dynamics)	Order rate	Independent	• Tight capacity constraints (in relation to the mean demand) result in high inventory shortfalls. • Increasing capacity is a necessary solution for agile manufacturing.
[xv] Ponte et al. (2017)	Simulation (differential equations modelling)	Order rate	Independent	• When capacity reduces, order variability decreases at the expense of an increase in inventory variability —and hence a reduced fill rate. • Overall, capacity limitations stop unnecessarily large orders being issued and this has some economic value.
[xvi] Shukla and Naim (2017)	Simulation (system dynamics)	Transportation and production	Independent	• Detecting capacity constraints is essential to improve the dynamic performance of supply chains.
[xvii] Framinan (2017)	Mathematical analysis	Order rate Lead time	Independent Load-dependent	• If capacity refers to the rejection of orders in excess of a threshold, the effect of capacity in the Bullwhip Effect is to dampen it. • If capacity constraints induce some variability in the lead times, the effect in the Bullwhip Effect is linked to the way the lead times and the demand are forecast

improvement was explained in subsequent works as a reduction in the variability of the orders issued throughout the supply chain (Wilson, 2007; Cannella et al., 2008; Ciancimino et al., 2009; Chen and Lee, 2012; Spiegler and Naim, 2014; Ponte et al., 2017), which clearly has an economic value for the members of this system. Interestingly, Hamdouch (2011) underlined that capacity limits also impact on market response through the selling price of products, which decreases the accuracy of demand forecasting —a second route through which capacity constraints impact on the generation of the Bullwhip Effect.

However, such decrease in order variability was found to be achieved at the expense of an increase in inventory variability (Vlachos and Tagaras, 2001; Ciancimino and Cannella, 2009; Nepal et al., 2012; Ponte et al., 2017). A direct consequence is that customer satisfaction tends to decrease as capacity tightens (de Souza et al., 2000; Wilson, 2007; Boute et al., 2009; Spiegler and Naim, 2014; Hussain et al., 2016; Ponte et al., 2017). In addition, a negative impact of capacity constraints on the agility of the supply chain was observed by Helo (2000) and Hussain et al. (2016).

Although the previous insights represent the main stream of research, some works reach different conclusions. For example, both Cannella et al. (2008) and Juntunen and Juga (2009) show that an increased capacity does not always result in an improved customer service level, while Nepal et al. (2002) do not observe a significant impact of capacity constraints on order variability. These contributions illustrate that the impact of capacity restrictions on supply chain performance heavily depends on the assumptions made. All in all, the aforementioned works highlight the significant impact of capacity constraints on the performance of supply chains, which cannot be ignored. In this sense, the overall system must be analyzed and these constraints must be detected (Shukla and Naim, 2017).

Table 1 also underscores the fact that the majority of these studies (14 out of 16) consider the lead time to be an independent parameter. This simplified approach to model the lead time is by far the most common assumption in the literature on supply chain dynamics (see *e.g.* Lee et al. 1997, Chen et al. 2000, and Dejonckheere et al. 2003). Indeed, as mentioned in Section 1, we have found only three studies dealing capacity constraint supply chain and load-dependent lead time, *i.e.*, Helo et al. (2000), Boute et al. (2009) and Framinan (2017). Helo (2000) concludes that idle capacity is not always non-productive: In this sense, he highlights that "cost efficiency and fast delivery are trade-off performances which cannot be maximized at the same time". Interestingly, Helo (2000), on the basis of some simulation and previous seminal works (see e.g. Chen et al. 1992, Burbidge 1996, Hernandez and Vollmer 1998) suggests that increasing the flexibility of capacity (the ability to change load capacity and operates on a master production schedule level and thus includes manpower and machinery) can improve supply chain performance. On the other hand, Boute et al. (2009) show how limits in production capacity tend to inflate the retailer's inventory requirements, which increases the costs of the supply chain. From this perspective, they explore the relationship between the proportional controller of the order-up-to model and the efficiency of the supply chain. Finally, Framinan (2017) suggests that if capacity constraints induce some variability in the lead times, the effect in the Bullwhip Effect is linked to the way the lead times and the demand are forecast.

All three studies have been developed under a number of rather restrictive assumptions and thus, the impact of capacity has been studied for limited market and decision-making scenarios. Therefore, our research is aimed towards enhancing the understanding of capacitated supply chain by modeling and exploring these scenarios." In the next section, we describe the supply chain model that we have considered.

3. Capacitated Supply Chain Model

In this section, we first describe how the capacity constraint can be modelled using a CT-TP curve. Then, we present a detailed description of the supply chain model, including definitions of the operational aspects. A summary of the notation employed is provided in Table 2.

3.1. CT-TP curve and capacity constraint.

In order to infer the effect of production capacity on supply chain performance, we assume that there is a causal relationship between orders release, capacity saturation and lead time. This relationship is theoretically known by Little's law, and it has been well-documented in empirical observations of lead times in real-life manufacturing and transportation systems (Upasani and Uzsoy, 2008; Fransoo and Lee, 2013; Kacar et al., 2016), where lead times strongly increase as the number of items in the pipeline reaches the maximum capacity of the production line (Yang et

Table 2. Notation of the supply chain model.

Variables			
O_t	replenishment order quantity the end of period t	TI_t	target inventory at the end of period t
d_t	market demand in period t	TW_t	target work in progress at the end of period t
$\hat{d}_t$	market demand forecast at the end of period t	W_t	work in progress at the end of period t
I_t	inventory of finished materials at the end of period t	L_t	lead time at the end of period t
$\hat{L}_t$	lead time forecast at the end of period t	Th_t	throughput at the end of period t
d^*_t	demand fulfilled in period t		
Parameters and Statistics			
ψ	capacity saturation	ε	safety stock factor
η	responsiveness factor	φ	capacity saturation factor
δ	stationary lead time	β	proportional controller
μ_d	mean of market demand	θ	lead time forecast smoothing factor
α	demand forecast smoothing factor	t	generic instant of time
σ_I^2	variance of inventory	σ_d/μ_d	coefficient of variation of the customer demand
σ_d^2	variance of market demand	σ_O^2	variance of order quantity

al., 2007). This behavior is usually captured using the so-called cycle time-throughput (CT–TP) curves (Fromm, 1992; Brown, 1997; Ankenman et al., 2011; Park et al. 2002), which empirically quantify the relationship between the average cycle time (ACT) (*i.e.*, the time an item takes to traverse the system, see Little, 1961) and the throughput rate (Yang et al., 2007). Figure 1 is a sample CT-TP curve adapted from Mönch et al. (2013), where it can be seen that the lead time changes depending on the work in progress. If the system is operating at the level of 22,000 units, by ramping up an additional 500 units, it experiences only a minor change in average cycle time. However, if the system is operating at the level of 22,500 units, a relatively low increase in the work in progress dramatically alters the lead time Generally, CP-TP curves exhibit a "hockey stick" shape (see *e.g.* Park et al., 2002; Fowler et al., 2001; Yang et al., 2007; Pahl et al., 2007; Kacar et al. 2016, among others). That is to say, lead times are almost stable for all orders up to a certain level of the production line, and increase exponentially when the system is saturated with excessive orders.

CT–TP curves are often employed in industry as decision-making tools in manufacturing settings (see Ankenman et al., 2011 and references herein) as they allow for the estimation of cycle time, which enables companies to make better capacity decisions. Literature in this area mainly focus on developing and improving methods for generating CT-TP curves (see *e.g.* Nemoto et al., 2000; Park et al., 2002; Fowler et al., 2001; Leach et al., 2005; Yang et al., 2007; Pahl et al., 2007; Veeger et al., 2010; Kacar et al., 2016, among others), given that the generation of empirical CT-TP curves requires collecting large amounts of representative data (Mönch et al., 2013).

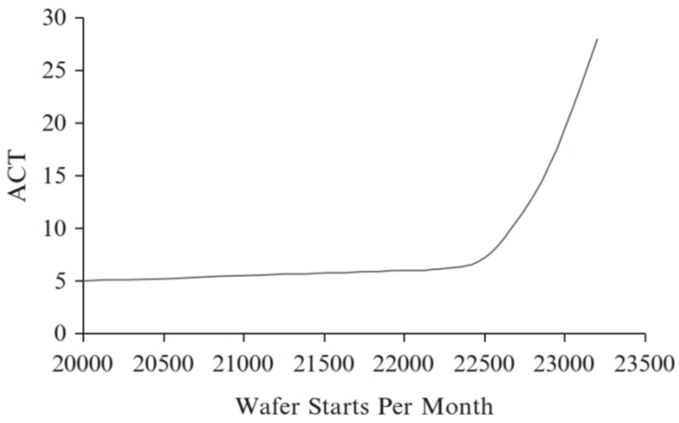

Figure 1. Empirical estimation of a CT-TP curve (source: Monch et al., 2013).

In the context of our research, modelling lead times using a CT-TP curve may represent a reasonable procedure to come closer to the true essence of capacity problems in real-life supply chains. Thus, we reproduce the CT-TP curve (Fig. 2) through a nonlinear analytical expression composed by two areas (Eq. 1).

$$L_t = \begin{cases} \delta & W_t < \psi \\ \delta e^{\frac{1}{\eta}\left(\frac{W_t}{\psi}-1\right)} & W_t > \psi \end{cases} \tag{1}$$

The quasi-horizontal state region in the CT-TP curve is modelled using a constant function $L_t=\delta$. When work in progress is lower than a saturation limit ψ (*i.e.*, the maximum number of items in work in progress that can be processed without saturating the production and distribution system), orders are fulfilled in a constant lead time (*i.e.*, the workload does not affect the lead times). Essentially, this region models the situation where the shop floor capacity can handle the current workload and therefore the lead time remains constant. Previous related studies (Evans and Naim, 1994; Simchi-Levi and Zhao, 2003; Cannella et al., 2014) modelling the capacity constraint condition as a limitation in the order quantity, have assumed that ψ is a function of the final marketplace mean demand μ_d and of a parameter named *capacity factor* φ, *i.e.*, $\psi = \mu_d \varphi$. In this study, as we are directly considering the saturation of the work in progress, we model ψ as a function of the final marketplace mean demand μ_d over the stationary lead time and the capacity factor $\varphi, i.e. \psi = \mu_d \delta \varphi$. In this manner, we are reproducing a real condition of the production-distribution system, in which the capacity is not dimensioned on the basis of the mean demand for each period, but on the basis of the mean demand during the expected lead time demand $\mu_d \delta$. The capacity factor φ may assume values within the range [1, ∞), where φ=1 indicates that the manufacturing capacity is equal to the mean demand multiplied by the constant lead time $\delta \mu_d$, while φ=∞ models the infinite production capacity.

The steep region of the curve (*i.e.*, when the work in progress exceeds ψ) is modelled as an exponential function depending on the steady state lead time δ, the work in progress W_t, the maximum capacity ψ, and a parameter named responsiveness factor η, which defines the slope of the curve. In this region the workload is assumed to affect the length of lead times, that is, it represents the case in which the current workload exceeds the capacity of the shop floor and the standard lead time cannot be guaranteed. Instead, a higher lead time is set, depending on the company's ability to absorb the current workload, *i.e.* to change its output to adapt to the new workload. We argue that this ability, also known in literature as volume responsiveness (Slack 1987, Holweg 2005) is captured by the parameter η, which controls the lead time increase if the work in progress exceeds ψ. This parameter can take values within the range [0, ∞). If η=0, we assume that the company is not able to deal with a workload exceeding the saturation level and the lead time would be infinite. As η increases from 0 to ∞, the slope of the CT-TP curve, according to which the lead time is decreasing in the work in progress, decreases.

To better define how η can capture the responsiveness of firms, we refer to the classification framework on responsiveness proposed by Reichhart and Holweg (2007). According to the authors, there are different types of responsiveness, both in terms of the unit of change (product, volume, mix and delivery responsiveness) and in terms of the time horizon affected (short, medium or even long-term responsiveness). Specifically, η suitably emulates the volume responsiveness, *i.e,* the ability of the system to change its output as a response to demand changes (see *e.g.*, Slack, 1987; Holweg, 2005; Reichhart et al., 2007; Reichhart and Holweg, 2007; Mahapatra et al. 2012; Bortolotti et al., 2013). In fact, the lower the volume responsiveness of the company ($\eta \rightarrow 0$), the worse the company is able to change its output to match the demand. Note that in the limiting case $\eta = \infty$ (perfect responsiveness), the company would be perfectly able to match any demand and therefore the lead times remain constant no matter the workload. In this sense, by modelling η we may capture (at a high level) factors that affect the company's ability to change (in the short-term) its output in response to customers' changes in the demand volume. In real-life supply chains, there are a number of factors influencing this ability. A detailed discussion of the options that companies may use to increase their responsiveness is given in Reichhart and Holweg (2007), and they include investments in flexible manufacturing, relying on higher inventory levels, changing the architecture of the product, or demand anticipation, among others. For instance, companies with flexible machinery and human workers (emulated in this study by high values of η) can adapt better to the turbulence of the market demand (and keep their standard lead times even when facing some demand peak) than those with less flexible resources (emulated by low values of η).

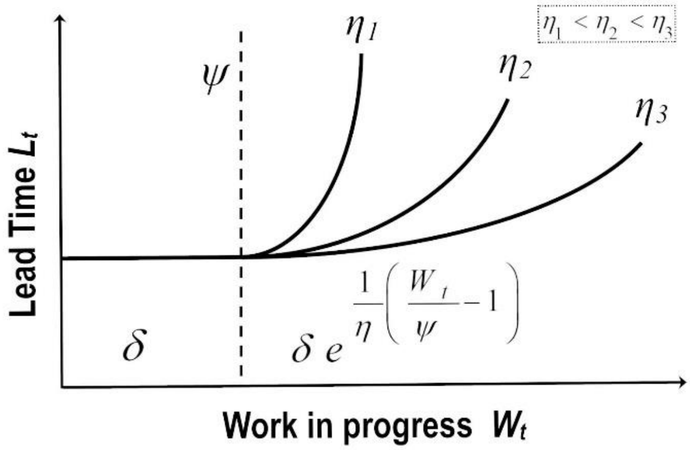

Figure 2. Iconic version of CT-TP curves

3.2 Supply chain model.

We adopt a methodology based on exploring the dynamics of the system (Riddalls et al., 2000) via a discrete time difference equation model (Holweg and Disney, 2005). This model has been implemented in Matlab R2014b, which allows us to fully automate the process of running the simulations, and hence enables exploring wide parametric designs in a time-efficient manner.

To generate a model that come closer to the true characteristics of real-life supply chains, we make assumptions based on insights from both axiomatic and empirical researches. Among these are the following:

- We focus on a single echelon of the supply chain, *i.e.,* the retailer, who satisfies the demand of a group of customers by ordering the product to a manufacturer. In this sense, we consider the relationships between three different nodes of the supply chain. This approach allows us to gain a deep understanding on the dynamic behavior of the capacitated supply chain, and thus to derive more specific managerial implications (see e.g., Chen et al., 2000; Wang and Disney, 2016; Naim et al., 2017, among many others).
- Backlogging is allowed as a consequence of stockholding (see *e.g.* Sterman, 1989; Udenio et al., 2015; Sterman and Dogan, 2015; Hussain et al., 2016). The backlog is fulfilled as soon as on-hand inventory becomes available.
- Non-negative condition of the order quantity. Recently, Chatfield and Pritchard (2013) analyse the impact of the allowance/disallowance of the return of goods on the Bullwhip Effect in a four-echelon serial supply chain. They show that allowing returns may result in a significantly larger Bullwhip Effect. Furthermore, the increase in order variance due to the returns may be quite dramatic at the upper echelons. Overall, their investigation of the impact of returns on the Bullwhip Effect question the default assumption (practically universal in Bullwhip Effect modelling), that returns are permitted. Motivated by the work of Chatfield and Pritchard (2013), and reasserted by Dominguez et al. (2015a) for a divergent supply chain, we assume that products delivered cannot be returned to the supplier. In this manner, we adopt a more reliable modelling assumption according to real-life supply chains.
- The exponential smoothing (Makridakis et al., 1982; Disney and Lambrecht, 2008) is adopted as forecasting method for estimating demand (α) and lead time (θ).
- Orders are generated according to two well-known (and largely adopted in practice) periodic-review inventory control policies, *i.e.*, the classical order-up-to (OUT) (Hax and Chandea 1984) and the smoothing replenishment rule. The former generates orders in which the entire gaps between target and the current levels of on-hand inventory, as well as the gap between the target and the current levels of the pipeline inventory. The latter generates orders to recover only a fraction these gaps. The amount of the gaps to recover is regulated by

decision parameters known as proportional controller β. If β=1, the policy is equivalent to the classical OUT, whereas if $\beta < 1$, we are emulating a smoothing replenishment rule. As β decreases, so the amount of the gaps between target and currents levels of inventory and work in progress do. The rationale for the smoothing order policy is to limit the tiers' over-reaction/under-reaction for changes in demand as orders are essentially "smoothed". Thus, by varying β we can explore if shifting from a classical OUT to a smoothing replenishment rule may impact the performance of the analyzed capacity constraints supply chain.

In each time period, the producer performs the following sequence of actions:

i. Receive ordered/processed products Th_t (Eq. (2)). These receptions correspond to orders placed several periods ago whose lead time depend on past work in process according to Eq. (3).

$$Th_t = W_{t-1} / L_{t-1} \tag{2}$$

$$W_t = W_{t-1} + O_{t-1} - Th_t \tag{3}$$

The throughput/receipt process is modelled according to Little's law (1968) (Eq. (2)) according to other relevant studies dealing with the dynamics of supply chains (see e.g., Sterman, 2000; Wikner, 2003; Deif and ElMaraghy, 2007; Chaudhari, 2011, among others). Equation (3) states that the work in progress W_t at time t is increased by O_{t-1} the order placed by the producer at time t-1 and is decreased by Th_t the throughput received at time t.

ii. Receive and satisfy the customer demand d_t. The fulfilled (or satisfied) demand d^*_t (Eq. (4)) can be obtained as the minimum value between the inventory available (i.e. the sum of the inventory in the previous period and the receipts) and the demand, as long as this is greater than zero.

$$d^*_t = \max\left\{\min\left(I_{t-1} + Th_t, d_t\right), 0\right\} \tag{4}$$

iii. Update the inventory of final products I_t (Eq. (5)). Note that the inventory position at time t is increased by the throughput at time t and decreased by the quantity d_t sent to the final customer at time t. Thus, I_t can be positive, representing storage or products at the end of the period, or negative, representing stock-outs.

$$I_t = I_{t-1} + Th_t - d_t \tag{5}$$

iv. Compute forecast $\hat{d}_t$ of the future demand (Eq. (5)) and a lead time forecast $\hat{L}_t$ (Eq. (6)).

$$\hat{d}_t = \alpha d_t + (1-\alpha)\ \hat{d}_{t-1} \tag{6}$$

$$\hat{L}_t = \theta L_t + (1-\theta)\ \hat{L}_{t-1} \tag{7}$$

Equations (6) models the exponential smoothing demand forecasting rule adopted for computing the Order O_t and the target inventory TI_t, and Eq. (7) models the exponential smoothing lead time forecast rule adopted for computing the the target work in progress TW_t. The value of α and θ reflects the weight given to the most recent observations.

v. Update the target inventory TI_t (Eq. (8)) and the target work in progress TW_t (Eq. (9)), respectively.

$$TI_t = \varepsilon \hat{d}_t \tag{8}$$

$$TW_t = \hat{L}_t \hat{d}_t \tag{9}$$

In Equation (7) TI_t the target inventory at time t is expressed as the product of the forecast demand of final customers $\hat{d}_t$ at time t times the safety stock factor ε, also known in literature as time to cover inventory (Sterman, 2000). In Equation (8), the target work in progress TW_t at time t is computed as the product of the demand forecast $\hat{d}_t$ and the lead time forecast $\hat{L}_t$ at time t.

vi. Place a replenishment/production order O_t (Eq. (10)).

$$O_t = \max\left\{\hat{d}_t + \beta\left(TW_t - W_t + TI_t - I_t\right); 0\right\} \tag{10}$$

Equation (10) models both classical OUT (if β=1) and the smoothing replenishment rule (if 0<β<1) (Disney and Lambrecht 2008). The order quantity is the sum of three components: (I) the forecast on the order from customer $\hat{d}_t$, (II) a fraction of the work in progress gap $\beta(TW_t - W_t)$, and (III) the inventory gap $\beta(TI_t - I_t)$. Finally, the logical operator "*max*" models the non-negative condition of order quantity.

4. Experiments, numerical results and analysis.

In this section, we first describe the experimental design by defining the factors to be analysed, the experimental outcomes (dependent variables), and the parameters of the model and simulation conditions. Then, we present a statistical analysis of the output data using ANOVA in order to assess the significance of the impact of the factors on the dependent variables.

4.1. Experimental design.

The main factor considered in the analysis is the capacity saturation factor (φ), although we also include five additional factors to study the impact of the capacity saturation in different scenarios: one related to the responsiveness of the manufacturing system (*i.e.*, the responsiveness factor η), two related to the order policy setting (*i.e.,* the safety stock factor, ε, and the proportional controller, β), one related to the lead time forecasting method (*i.e.*, the lead time forecast smoothing factor, θ), and a exogenous factor related to customer demand (*i.e.,* the coefficient of variation of customer demand, σ_d/μ_d). Since the main factor under analysis is φ, we select six levels for this factor and three levels for the other factors, as shown in Table 3. We adopt a full factorial design, so in total we perform 1458 different experiments.

To set up the customer demand we assume that the quantities ordered represent a standard product with repetitive demand and thus, we use a normal distribution (see *e.g.*, Chatfield 2013). To generate a more realistic representation of the production-distribution system, we assume that, when capacity saturation has been largely exceeded (*i.e,* work in progress is very high), lead-time cannot exceed by four times the stationary lead time value *(δ)* (see Table 3).

The dependent variables are three common metrics assessing the operational performance and the customer service level. More specifically, the operational performance is measured using the Order Rate Variance Ratio (*ORVrR*) and the Inventory Variance Ratio (*IVrR*), while the customer service level is measured via the Fill Rate (*FR*). The reduction of *ORVrR* and *IVrR* reflects improved cost effectiveness of supply chain members’ operations, while the increase of FR indicates lower stock-out costs.

Table 3. Experimental design.

Model parameters	*Value*					
Customer demand (d)	$N(\mu_d = 100, \sigma^2)$ units per period					
Stationary lead time (δ)	2 periods					
Demand smoothing forecasting factor (α)	0.2					
Independent variables	*Values*					
Capacity saturation factor (φ)	1.25	1.50	1.75	2.00	2.25	2.50
Responsiveness factor (η)	1.0		2.0		4.0	
Customer demand variability (σ_d/μ_d)	0.15		0.30		0.45	
Lead time smoothing forecast factor (θ)	0.1		0.5		1.0	
Safety stock factor (ε)	0.5		1.0		3.0	
Proportional controller (β)	0.3		0.6		1	
Dependent variables						
Order Rate Variance Ratio (ORVrR)						
Inventory Variance Ratio (IVrR)						
Fill Rate (FR)						

$ORVrR$ was proposed by Chen et al. (2000) and is the most common demand amplification measure in the literature (Disney and Lambrecht, 2008). It is defined as the ratio between the demand variance at the downstream and at the upstream stages (Miragliotta, 2006) (Eq. 13).

$$ORVrR = \frac{\sigma_O^2}{\sigma_d^2} \tag{11}$$

$IVrR$ was proposed by Disney and Towill (2003) to measure net stock instability, as it quantifies the fluctuations in actual inventory against the fluctuation in demand, see (Eq. 14). An increased inventory variability results in higher holding and backlog costs, and increasing average inventory costs per period (Disney and Lambrecht, 2008).

$$IVrR = \frac{\sigma_I^2}{\sigma_d^2} \tag{12}$$

The Fill Rate (Eq. 17) is a key metric of customer satisfaction within supply chains, which measures the demand fulfilment experienced by the consumer. It can be obtained as the ratio of the mean fulfilled demand to the mean demand, see e.g. Ponte et al. (2017).

$$FR = \frac{d^*_t}{d_t}\% \qquad (13)$$

The length of simulations runs is set to T=20,000 periods, ensuring reaching a steady state in the system, and data from the first 200 periods of each replication are removed as warm-up. In order to account for randomness and test the statistical significance of the experimental factors, we perform several replications of each experiment. According to Kelton et al. (2007), when the half-width of confidence interval is smaller than a user-specified value (*e.g.* within 10% of the mean, Yang et al. 2011), the number of replications is acceptable for statistical analysis. Due to the high length of the simulation runs, the obtained output randomness is very low, and we achieve this condition with a low number of replications (i.e. 5 replications). Finally, in order to be able to compare the different replications, we fix the random number generator, and thus the differences obtained are due to the parameter settings and not randomness.

4.2. Analysis of results.

The statistical analysis of the simulation data is carried out via an Analysis of Variance (ANOVA), with a level of significance of p=0.05. Normality of data output was checked through Kolmogorov-Smirnov and Shapiro-Wilk tests. Table 4 shows the results of the ANOVA for the three performance metrics. Due to the numerous interactions among factors that result from the experimental design, and since the main factor in the scope of this analysis is the capacity factor φ, we show only the (first-order) interactions that involve this factor and do not focus on all other interactions from our analysis. Results show that all factors and interactions are statistically significant at 95% confidence level, so we reject the null hypothesis that there is no difference in means between the different groups. In the following we analyse the results using main effects and interaction plots together with the results obtained from ANOVA.

We start our analysis by focusing on the main effects of φ on *ORVrR*, *IVrR*, and *FR*. Looking into Table 4 (F-value) it can be noticed that φ and η have a similar impact on $ORVrR$ and $IVrR$, being φ statistically more significant than η in FR. φ is also more significant than σ_d/μ_d for $ORVrR$ and $IVrR$, while for FR the opposite happens. The impact of φ on all the three metrics is less significant when compared with ε and β (*i.e.*, ε and β have a direct impact on how orders are placed and thus, they have a big impact on all the three metrics). Finally, θ has the lowest impact on all the three metrics.

Table 4. ANOVA for *ORVrR*, *IVrR* and *FR*.

	ORVrR			***IVrR***			***FR***		
Source	*DF*	*F*-value	*p*	*DF*	*F*-value	*p*	*DF*	*F*-value	*p*
φ	5	1392,771	<0.001	5	1375,480	<0.001	5	511,604	<0.001
η	2	1511,993	<0.001	2	1456,671	<0.001	2	408,381	<0.001
σ_d/μ_d	2	1246,008	<0.001	2	742,976	<0.001	2	87046,411	<0.001
ε	2	21028,743	<0.001	2	23935,971	<0.001	2	90666,789	<0.001
β	2	133877,414	<0.001	2	3310,967	<0.001	2	1956,753	<0.001
θ	2	153,731	<0.001	2	53,489	<0.001	2	36,971	<0.001
$\varphi * \beta$	10	206,741	<0.001	10	5,025	<0.001	10	1,124	0.339
$\varphi * \varepsilon$	10	24,641	<0.001	10	3,345	<0.001	10	125,721	<0.001
$\varphi * \eta$	10	1140,242	<0.001	10	953,171	<0.001	10	292,303	<0.001
$\varphi * \theta$	10	87,530	<0.001	10	70,620	<0.001	10	27,125	<0.001
$\varphi * \sigma_d/\mu_d$	10	143,932	<0.001	10	425,874	<0.001	10	242,198	<0.001
	Adjusted R^2 (%) = 98.1			Adjusted R^2 (%) = 92.6			Adjusted R^2 (%) = 98.6		

Figure 4 shows the main effects of φ and confidence intervals at a 95% confidence level. It is observed that the performance of the system improves significantly if the capacity is high as compared to scenarios with tight capacity restrictions (*i.e.* increasing φ reduces *ORVrR* and *IVrR* and increases *FR*). This result implies that, if the capacity of the system is large, lead times tend to stabilize around the minimum stationary value, thus improving the performance of the supply chain. On the contrary, when there is a tight capacity limit, the system works under saturation and lead times tend to increase, which worsens the supply chain response. Interestingly, the supply chain performance stabilizes for a certain value of φ (*i.e.*, additional increments in φ do not produce any performance improvement). This output suggests that, if the capacity becomes large enough, the capacity constrained system turns into an unconstrained supply chain. In the literature, such supply chains are known as the so-called exogenous supply chains (Hum and Parlar, 2014), in which lead times are independent from internal variables. Under the boundary conditions of our simulation analysis, *ORVrR, IVrR* and *FR* stabilize, in average, for φ~1.75 (when the capacity is around 75% higher than the mean demand during lead times).

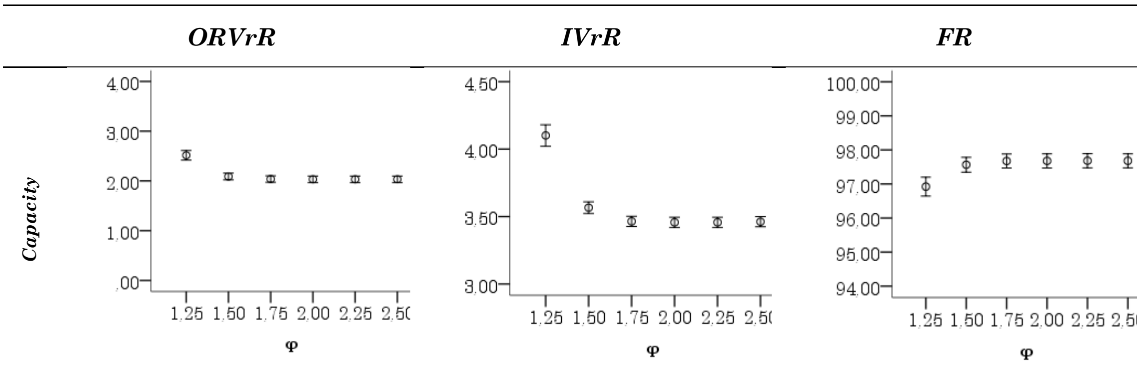

Figure 4. Main effects and 95% confidence intervals (φ).

Figure 5 shows the main effects of the other experimental factors (*i.e., η,* σ_d/μ_d, ε, β and θ) and confidence intervals at a 95% confidence level. Looking at the responsiveness factor η, a reduction of *ORVrR* and *IVrR* is obtained when η increases from η=1 to η=2, while no significant change can be observed when η increases from η=2 to η=4. In addition, an increase of η improves FR. Thus, by increasing this factor (*i.e.* increasing firm's capability to respond to variations in the market), the system is able to mitigate the variance of the orders received, which translates into more stable production schedules and inventories and higher customer service levels. In other words, the dynamics of the system dramatically suffer from not being able to deal with high orders in a time-effective manner, since lead times are longer. However, the benefit of increasing the system's responsiveness is limited, since once the system has reached enough responsiveness to handle demand variations, it is no longer benefited from higher responsiveness.

In Figure 5, we also show the main effects of σ_d/μ_d, ε, β and θ in order to provide a general picture of the simulation outputs. Note that impact of some of these factors has already been analysed in the literature (see e.g. Chatfield et al., 2004: Disney and Lambrecht, 2008; Ciancimino et al., 2012; Syntetos et al. 2011, among others). In this research, we are more interested in the interactions between these factors and the capacity saturation factor φ, which will be analysed later (see Table 4 and Figure 6). Nonetheless, we would like to remark that the results regarding the impact of these parameters on supply chain performance are in line with observations showed by previous related empirical and theoretical studies. This concordance supports the validation of our model.

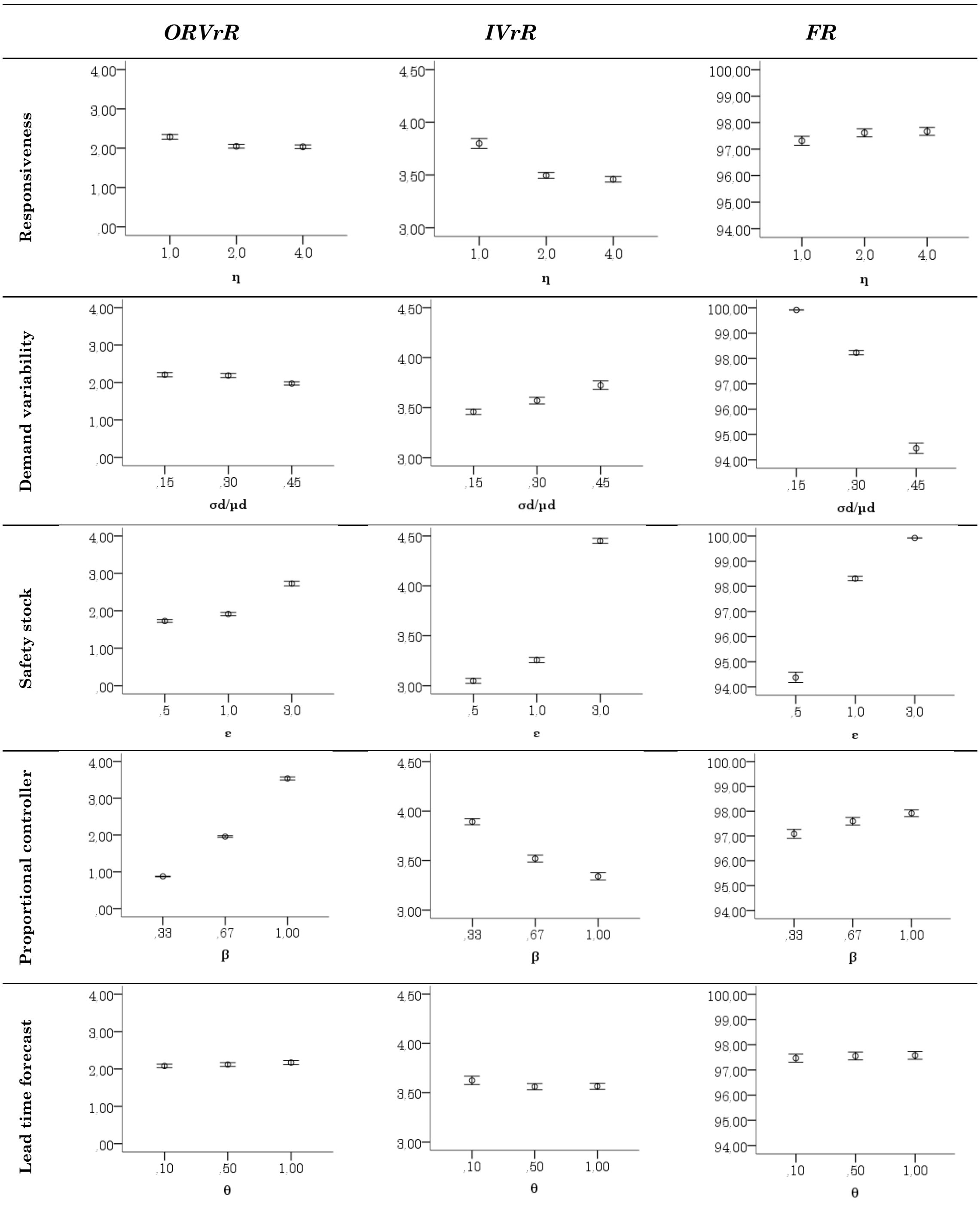

Figure 5. Main effects and 95% confidence intervals (η, σ_d/μ_d, ε and β).

The ANOVA reports significant interactions between φ and the other experimental factors (η, σ_d/μ_d, ε and β), as it can be seen in Table 4. Some of these interactions have an important impact on *ORVrR, IVrR* and *FR* (they show a high *F*-value) and, as such, the impact of φ on supply chain performance needs to be interpreted by considering the interactions with other factors (Figure 6). Overall, the interaction between φ and η is the most important one. When η is high (η=4), φ barely impacts on performance (only a small performance increase is observed by increasing φ from

φ=1.25 to φ=1.50). This happens because the system only achieves the capacity limit in a very low percentage of the time periods, which has a marginal impact on the overall results. Thus, a manufacturer with high responsiveness is able to avoid the amplification of the variability in the system, achieving the minimum *ORVrR* and *IVrR* or maximum *FR* with a low capacity (φ~1.50). On the contrary, as η decreases, the system becomes very sensitive to φ, and the capacity needed to achieve such minimum/maximum performance increases (for η=1 *ORVrR* and *FR* stabilizes at φ~1.75, and *IVrR* stabilizes at φ~2.00). It is also important to point out that a manufacturer operating with low capacity and low responsiveness benefits from a significant performance improvement by either increasing its capacity and/or its responsiveness.

The next interaction in importance takes place between φ and σ_d/μ_d, in particular for *IVrR* and *FR*. When σ_d/μ_d is low (σ_d/μ_d=0.15) φ barely impacts on *IVrR* and *FR*. However, as σ_d/μ_d increases, the supply chain becomes more sensitive to φ, and the capacity needed to achieve the minimum *IVrR* and maximum *FR* increases. More specifically, the capacity needed to achieve the minimum *IVrR* (maximum *FR*) is φ~1.75 (φ~1.50) when σ_d/μ_d=0.30 and φ~2.00 (φ~1.75) when σ_d/μ_d=0.45. Therefore, higher variability of demand requires higher capacity to avoid performance deterioration. Finally, *ORVrR* improves when φ increases from φ=1.25 to φ=1.50 for all the analysed values of σ_d/μ_d. After that value, any increment of φ does not improve *ORVrR,* except if σ_d/μ_d=0.45, where *ORVrR* can be reduced by increasing φ from φ=1.50 to φ=1.75. To sum up, a supply chain operating in a market characterised by high variability is more vulnerable to capacity limitations.

The interaction between φ and ε is particularly important for the *FR*. If the safety stock factor is set to high value (ε=3.0) the system maintains a high stock level, which is enough to handle the variability of market demand and thus, a high *FR* can be obtained regardless the capacity of the manufacturer. As ε decreases, the system maintains lower stock levels and becomes more sensitive to the value of φ (*i.e.* the system needs a higher capacity to react to the incoming orders). Specifically, for low values of ε (ε=0.5, ε=1.0) the *FR* decreases when φ is below φ~1.75. In case of *ORVrR* and *IVrR*, performance deterioration occurs for any value of ε if φ is below a certain value (φ~1.50 and φ~1.75, respectively). However, this deterioration may be slightly different depending on the value of ε. To sum up, the customer service level of a supply chain characterised by relevant capacity restriction strongly depend from the safety stock factor.

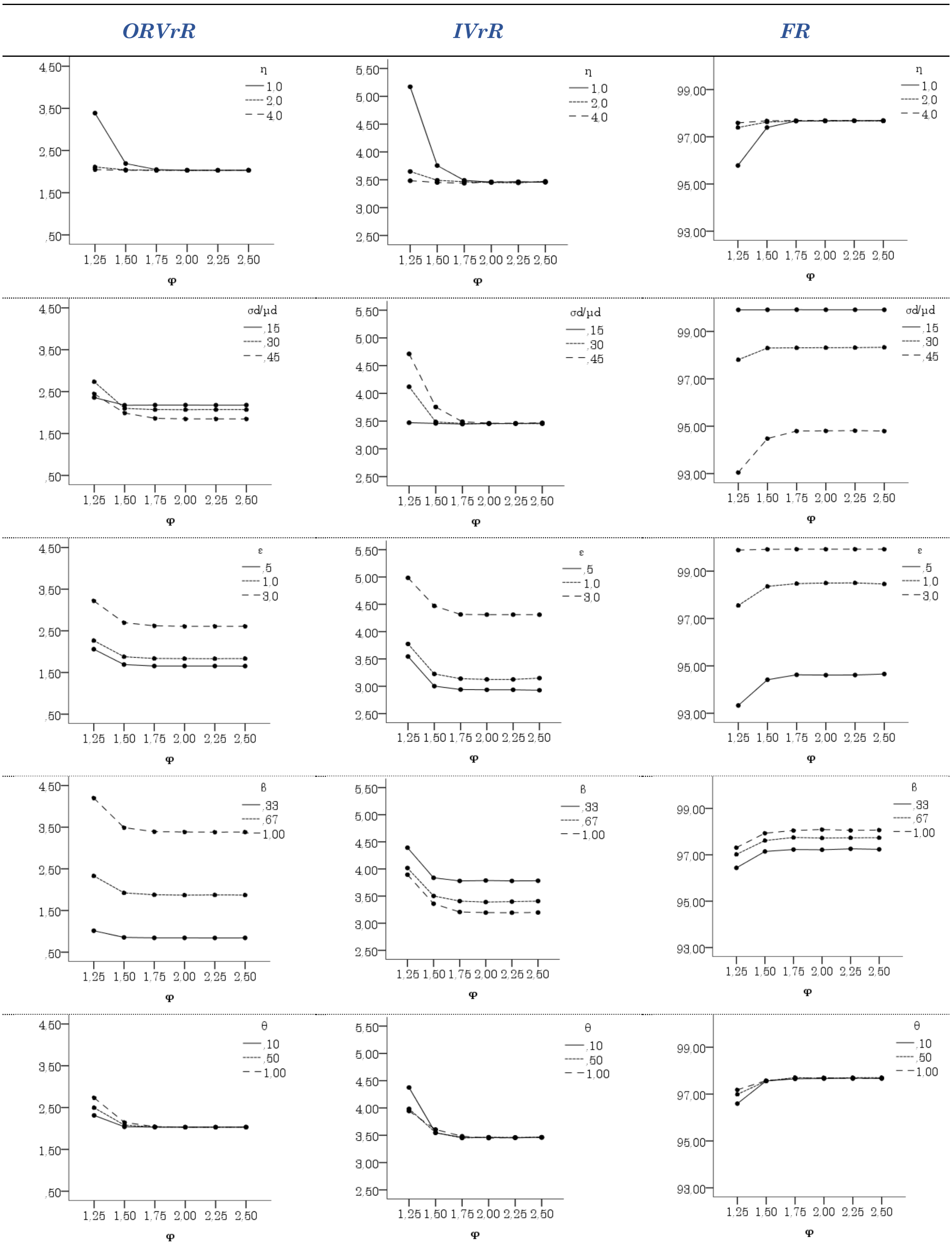

Figure 6. Interaction plots with capacity saturation factor (φ).

Regarding the interaction between φ and β, the performance deterioration (*ORVrR* and *IVrR*) caused by a low capacity is higher for a classical OUT (β=1) than for a smoothing replenishment rule (β=0.67 and β=0.33). In fact, the smoothing replenishment rule homogenises production orders (by filtering the incoming demand), making the current manufacturing capacity more efficient. This is particularly important for very low capacity settings (φ=1.25), where we observe the maximum *ORVrR* improvement obtained from the implementation of the smoothing replenishment rule. On the contrary, the deterioration of the FR observed for low capacity settings is very similar for both policies. To sum up, the adoption of a smoothing replenishment rule in a supply chain characterised by relevant capacity restriction may strongly reduce the demand amplification phenomenon.

The last interaction under analysis takes place between φ and θ. This interaction is significant only when the producer operates with low capacity (φ<1.50). In this scenario, the forecasting parameter has a significant impact on performance. Adopting a low θ (*i.e.* a low-reacting lead time forecast that slowly adapts to the changes in lead times) smooths *ORVrR* but deteriorates *IVrR* and *FR*. On the contrary, adopting a higher θ (*i.e.* a highly-reactive lead time forecast, which relies more on recent lead times to estimate future lead times) improves *IVrR* and *FR* but has a negative impact on *ORVrR*. In this sense, we can conclude that intermediate values of θ (*e.g.*, θ=0.5) provide the better trade-off between production variability (*ORVrR*) and inventory performance (*IVrR*, *FR*). On the contrary, it should be noted that, if the producer operates with high capacity (φ≥1.50), the forecasting parameter θ has no meaningful impact on performance. To sum up, the importance of adjusting appropriately the forecasting method in order to balance the trade-off between order stability and inventory performance grows when the capacity of the manufacturer diminishes.

4.3. Summary of findings

As a general conclusion, our results show that, in a capacitated supply chain with load-dependent lead time, the capacity limitation presents a significant impact on supply chain performance and that this impact depends on the responsiveness of the supply chain, the market condition (demand variability) and the replenishment decisions (safety stock factor and proportional controller). More specifically we have identified the following six findings:

1. *As the producer's capacity increases (i.e., the manufacturer is able to deal with higher WIP without reaching the saturation, that is, maintaining a constant lead time), supply chain performance improves as well*. Assuming load-dependent lead times implies that, when the capacity of the manufacturer is low, the supply

chain experiments an abrupt increment of the mean and the variability of the replenishment lead time. As widely documented in the literature, this aspect significantly contributes to the amplification of order variance and, consequently, to the degeneration of supply chain performance. Our results reassert these evidences and highlight the obstacle created by capacity constraints, particularly in terms of Bullwhip Effect. In contrast with several works dealing with the dynamics of capacitated supply chains (see Table 1), this study shows that capacity constraints do not only negatively impact customer service level, but also in terms of operational efficiency. In fact, according to our modelling assumption, capacity does not act like a Bullwhip Effect damper. Contrarily, it may exasperate the Bullwhip Effect, increase inventory variability and cause stock-outs. This contrasting result may have remarkable implications for both supply chain researchers and practitioners (see subsections 5.1, 5.2 and 5.3).

2. *As the producer's responsiveness decreases (i.e., the lead time strongly increases once the capacity saturation of the system is reached), the dynamic performance of the supply chain rapidly degenerates*. Analogously to the previous consideration, a low responsiveness affects lead times and, consequently, supply chain performance. A high responsiveness factor implies that, if the WIP exceeds the maximum capacity (*i.e.*, the producer reaches the saturation), lead times do not increase as drastically, exponentially as for a low responsiveness factor. Contrarily, a low responsiveness factor implies an exponential boost of lead times caused by an excess of WIP and thus, a significant deterioration of operational efficiency and customer service level.

3. *The negative impact of low capacity on supply chain performance is exacerbated by a low responsiveness factor*. *On the contrary, a high responsiveness factor is able to soften the negative effects of a low capacity*. A system with high responsiveness is able to accommodate the excess of WIP caused by low capacity, maintaining short lead times and thus alleviating the detrimental consequences of high lead times. On the contrary, a system with low responsiveness needs a high capacity to limit the saturation (*i.e.,* excess of WIP) occurrences, which results in very long lead times.

4. *A supply chain operating in a market characterized by high variability is more vulnerable to capacity limitations in terms of inventory variance and customer*

service level. A high capacity is required to deal with orders which are considerably higher than the average, avoiding frequent exceeds of the capacity and consequently large and variable lead times. On the contrary, a market demand with a low variability allows to maintain a low capacity.

5. *The customer service level of a supply chain characterized by limited capacity strongly depends on the safety stock factor*. With a high safety stock factor the system maintains a high stock level, which allows to meet the incoming demand and thus, a high customer service level can be obtained regardless the current capacity. A low safety stock factor requires higher capacity to react to the incoming demand, since the level of stock is often not enough to cover such demand.

6. *The negative impact of a low capacity on the demand amplification phenomenon may be diminished by the adoption of a smoothing replenishment rule.* The smoothing replenishment rule filters the variability of incoming orders and produces more stable production orders, thus making possible to efficiently handle the demand with lower capacity.

7. *As the producer's capacity decreases, the lead time forecasting method impacts more significantly on supply chain performance.* When the producer operates with a high capacity, the adopted lead time forecasting factor does not impact on both operational performance and customer service level. However, for low capacity settings, the dynamics of the supply chain becomes very sensitive to the adopted lead time forecasting factor. Specifically, low-reactive lead time forecasts contribute to (slightly) decrease the Bullwhip Effect; however, this occurs at the expense of significantly worsening the inventory performance of the supply chain. Contrarily, highly reactive lead time forecasts produce better performance in terms of inventory costs and customer service level, but they may increase the variability of the orders.

5. Implications for research and industry

The six findings identified in the previous section suggest relevant implications for researchers and practitioners. In the following subsection we describe three possible implications regarding the modelling assumptions adopted for studying the supply chain dynamics and challenges for real-life supply chains investments.

5.1 Capacity constraints and supply chain analysis and modelling: Re-thinking assumptions

The assumption of unconstrained capacity has been largely accepted in supply chain dynamics and Bullwhip Effect studies. The assumption that the customers are fulfilled in a certain steady-state replenishment time regardless the order size facilitates the modelling and simulation efforts for the analysis of complex supply chain scenarios. However, this study shows that considering (or not) load-dependent lead times in the supply chain model may considerably alter the estimation of supply chain performance. Furthermore, by comparing findings of previous related works (see Table 1) with our results, we note that, depending on "how" the capacity constraint is modelled (*e.g.,* load-dependent lead time, limit in the orders placed, rejection of orders), similar simulation experiments (*i.e.,* identical supply chain structure and methodology) may produce dissimilar outputs. In general, studies modelling the capacity restriction through a limitation to the orders placed to suppliers or to the orders' acceptance channel have observed a reduction in terms of the Bullwhip Effect. Contrarily, modelling the capacity through load-dependent lead times suggests that capacity constraints can be an important cause of this phenomenon.

In this fashion, this work contributes to a relatively new research stream on Bullwhip Effect and supply chain dynamics aimed at improving the understanding of the impact of certain modeling assumptions on the results provided by classic supply chain models (see *e.g.*, Towill et al., 2007; Chatfield and Pritchard, 2013; Chatfield, 2013; Dominguez et al., 2015a,b). More specifically, we reassert the relevancy of modelling capacity constraint for supply chain and highlight the need for exploring, testing, and validating further reliable empirically-driven modelling assumptions of capacity in supply chains.

5.2 Capacity planning and lead time compression: Reflecting on the real-life supply chain investments

Capacity planning is essentially important for effective strategic decision-making in various industries (Xie et al., 2014). Companies invest for sufficient capacity to move and store its goods to meet demand in the next cycle of its activities (Crainic, 2016). The problems concerning capacity planning and allocation in the supply chain are challenging due to long production lead times and high demand uncertainty. From a supply chain dynamics viewpoint, the related literature has mainly shown that capacity constraints may act as a Bullwhip Effect limiter (see Table 1). Assuming a load-dependent lead time on the basis of empirical evidence, we show that high capacity improves the dynamics of supply chain and low capacity increases the Bullwhip Effect, inventory holding costs and stock-out costs. Thus, an intuitive implication should appoint to invest in additional capacity, also considering the detrimental effect produced by the Bullwhip Effect (Lee 2010). However, the cost of increasing the manufacturing capacity, as well as the overhead costs of maintaining such capacity, are considered fixed costs and they are excluded from the product cost (De Matta 2017). Thus, the underutilization of capacity creates several potential unnecessary costs. As stated by Disney and Lambrecht (2008), "companies have to invest in extra capacity to meet the highly variable demand. This capacity is then under-utilized when demand drops. Unit labor costs rise in periods of low demand, overtime, agency, and sub-contract costs rise in periods of high demand." In this context, it is relevant to find an appropriate trade-off between fixed costs due to large capacity and inefficiencies caused by the impact of demand variability on the saturation of the system.

In this fashion, our work provides further insights to improve the efficiency of this trade-off analysis. Our results suggest that the capacity planning should also consider the nonlinear dependence between workload and lead times, and thus, on how lead times increase when the system is over-utilized. A supply chain system characterized by high responsiveness (*e.g.*, flexible manufacturing, advanced demand information, etc.) would perform better when the system is saturated (Singh et al. 2014), as it will be able to limit the subsequent increase of production lead time by better absorbing the increase of demand. However, to achieve a high responsiveness, supply chains have to assume further variable and fixed costs for acquiring new technologies (*e.g.*, flexible machines), overworks, subcontracting etc. Thus, depending on both market/operational and sector condition costs, supply chains may adopt different capacity planning strategies. For instance, supply chains facing a stable and/or predictable demand may maintain a tight capacity (*i.e.*, close to the average demand) and a relatively low responsiveness. Contrarily, supply chains facing significant uncertainties and/or high variability of orders may opt among (1) maintaining an elevated capacity, avoiding the need of a high responsiveness, (2) maintaining a tight capacity and an elevated responsiveness, or (3) achieving a compromise between capacity and responsiveness.

Regardless the operating scenario and the variable and fixed sector costs, this study reasserts the need for investing in reducing and stabilizing lead times: The "lead time compression principle" (Towill, 1996) continues to be an *aere perennius* in operations management (Ciancimino et al., 2012), and particularly in the age of sustainability (Cannella et al. 2016) and turbulence (Christopher and Holweg, 2017), it is extremely relevant when supply chains tend to saturate their resources.

5.3 Fixed capacity and responsiveness settings: the role of market demand and replenishment decisions

Until now we have discussed the relationship between the capacity and the responsiveness of the production system and the trade-off existing between the two. We now address a scenario where both the capacity and the responsiveness of the production system are low and cannot be easily changed (e.g. a sudden increase in demand mean, which would require increasing the capacity to maintain the current performance, or a partial re-structuration of the production system, which may temporarily reduce the current capacity and/or responsiveness). In such scenario, demand uncertainty plays a crucial role, as high demand uncertainty requires a high capacity in order to control the inventory variability and to keep a high customer service level. Therefore, whereas the procurement cannot directly influence future demand, a closer collaboration among procurement, marketing and forecasting of the company may achieve the following goals: (1) to smooth the market demand by adopting specific marketing strategies (Klassen and Rohleder, 2002) and (2) to reduce demand forecasting errors. If the company prioritizes the customer service level, another action that can be taken is to increase the safety stock factor, since it prevents the fill rate reduction caused by the lack of sufficient capacity. Analogously, highly-reactive lead time forecasts (*i.e.*, strongly based on the most recent observations of lead times) can contribute to achieve higher customer service level and lower inventory holding costs, but they may also increase order variability. Finally, the adoption of smoothing replenishment rules and low-reactive lead time forecasting may mitigate the variance amplification phenomenon, particularly in the presence of scarce capacity.

6. Conclusions and further directions

In this paper we offer a novel perspective on capacitated supply chains by assuming load-dependent lead times. Based on empirical evidence, we model the manufacturing lead times as a nonlinear function depending on the current work in progress at the manufacturer and its capacity saturation limit and responsiveness. By doing so, we are able to move a step further in the understanding of the dynamic behaviour of capacitated supply chains by considering the response of the system after capacity saturation (*i.e.*, the order size cannot be fulfilled in the average stationary lead time). In this situation, the lead time of the order to be fulfilled depends on the current work in progress and the responsiveness of the manufacturing system.

In our analysis, we consider two factors to model the manufacturing lead time: the capacity saturation factor and the responsiveness factor. In order to determine how different manufacturing settings perform in different scenarios, we consider three additional factors, i.e. customer demand variability, safety stock factor, and proportional controller. Via discrete time difference equations approach, we perform a comprehensive simulation analysis to determine how these operational factors impact orders and inventory variability and customer service level. We show that, in contrast to other previous studies, increasing the capacity limit of the manufacturer has a positive effect on supply chain dynamics derived from maintaining a lower and constant lead time. On the contrary, reducing such capacity has a negative effect on performance. In the event of a manufacturer working under saturation (*i.e.*, the current work in progress is higher than its capacity), the responsiveness of the manufacturing system plays an important role (*e.g.*, a manufacturer with low responsiveness is not be able to efficiently manage the capacity saturation, resulting in long and variable lead times). We find a strong interaction between both factors, showing how the responsiveness of the manufacturer becomes more critical if she/he is working close to the capacity limit. Finally, we show that there are strong interactions among capacity, customer demand variability, safety stock factor, proportional controller and lead time forecasting factor. More specifically, we show that (1) a high demand variability, (2) a low safety stock factor, (3) a high proportional controller and (4) a highly reactive lead time forecast exacerbate the negative impact of a low capacity in terms of Bullwhip Effect.

The most important limitation of our work concerns the modelling assumption of load-dependent lead times. We assume that orders that do not exceed the capacity saturation limit are fulfilled in a constant lead time. However, modelling the capacity is still an issue, since all the complexities of a real manufacturing system cannot be captured by an implicit capacity model such as the CT-TP curve. Further studies should consider to explicitly modelling the capacity and/or to extend the current research by considering the stochasticity of lead-times and other methods for emulating the load-dependent lead time, *e.g.* clearing functions (Orcun, 2009; Mönch, 2012).

Another limitation concerns the supply chain topology. We have studied a two-node supply chain. However, this assumption is a simplification of real supply chains, which often show more complex structures (Chatfield et al., 2013; Dominguez et al., 2015a). Thus, the capacity-constrained supply chain needs to be analyzed in more realistic structures, such as divergent, convergent, conjoined, or the closed loop supply chain. Also, upstream members may limit/amplify the effects of both capacity and responsiveness of downstream members on supply chain dynamics, *e.g.*, how important is to increase the capacity/responsiveness of a downstream manufacturer if there is an upstream echelon with tight capacity/responsiveness? In addition, we note that we assumed specific boundary conditions and parameters for the experiments. Even if they have been selected and organized to provide general and reliable results, the impact of capacity in supply chain dynamics needs to be explored under other scenarios in order to contrast the obtained results (*e.g.* demand distributions, order policies, collaboration strategies, etc.). Finally, considering the impact of demand correlation on the link between capacity constraints and supply chain performance is also a line of research worth pursuing.

Acknowledgements

We are grateful to the anonymous referees and the editorial team as their suggestions allowed us to improve earlier versions of this manuscript. This research was supported by the Italian Ministry of Education, University and Research (Rita Levi Montalcini fellow), by the University of Seville (V PPIT-US), and by the Spanish Ministry of Science and Innovation, under the project PROMISE with reference DPI201680750P.

References

Ankenman, B. E., Bekki, J. M., Fowler, J., Mackulak, G. T., Nelson, B. L., Yang, F. (2011). Simulation in production planning: an overview with emphasis on recent developments in cycle time estimation. *Planning Production and Inventories in the Extended Enterprise*, 565-591, Springer.

Bortolotti, T., Danese, P., & Romano, P. (2013). Assessing the impact of just-in-time on operational performance at varying degrees of repetitiveness. *International Journal of Production Research*, 51(4), 1117-1130.

Boute, R. N., Disney, S. M., Lambrecht, M. R., Van Houdt, B. (2009). Designing replenishment rules in a two-echelon supply chain with a flexible or an inflexible capacity strategy. *International Journal of Production Economics*, 119(1), 187-198.

Boute, R., Lambrecht, M. (2007). Altruistic behavior in supply chain management. *Tijdschrift voor Economie en Management*, 52(3), 499.

Brown, S., Chance, F., Fowler, J. W., Robinson, J. (1997). A centralized approach to factory simulation. *Future Fab International*, 3, 83-86.

Buchmeister, B., Friscic, D., Palcic, I. (2014). Bullwhip effect study in a constrained supply chain. *Procedia Engineering*, 69, 63-71.

Burbidge, J. L. (1996). Period batch control (Vol. 12). Oxford University Press.

Cachon, G. P., Lariviere, M. A. (1999). Capacity choice and allocation: Strategic behavior and supply chain performance. *Management Science*, 45(8), 1091-1108.

Cannella, S., Bruccoleri, M., Framinan, J. M. (2016). Closed-loop supply chains: What reverse logistics factors influence performance? *International Journal of Production Economics*, 175, 35-49.

Cannella, S., Ciancimino, E., Marquez, A. C. (2008). Capacity constrained supply chains: A simulation study. *International Journal of Simulation and Process Modelling*, 4(2), 139-147.

Cannella, S., Framinan, J. M., Barbosa-Póvoa, A. (2014). An IT-enabled supply chain model: a simulation study. *International Journal of Systems Science*, 45(11), 2327-2341.

Chatfield D.C., Pritchard A.M. (2013). Returns and the bullwhip effect. *Transportation Research Part E: Logistics and Transportation Review*, 49 (1), 159–175.

Chatfield, D.C. (2013). Underestimating the bullwhip effect: a simulation study of the decomposability assumption. *International Journal of Production Research*, 51 (1), 230-244.

Chen, F., Drezner, Z., Ryan, J. K., Simchi-Levi, D. (2000). Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, lead times, and information. *Management Science*, 46(3), 436-443.

Chen, I. J., Calantone, R. J., & Chung, C. H. (1992). The marketing-manufacturing interface and manufacturing flexibility. Omega, 20(4), 431-443.

Chen, L., Lee, H. L. (2012). Bullwhip effect measurement and its implications. *Operations Research*, 60(4), 771-784.

Childerhouse, P., Disney, S. M., Towill, D. R. (2008). On the impact of order volatility in the European automotive sector. *International Journal of Production Economics*, 114(1), 2-13.

Christopher, M., Holweg, M. (2017). Supply chain 2.0 revisited: a framework for managing volatility-induced risk in the supply chain. *International Journal of Physical Distribution & Logistics Management*, 47(1), 2-17.

Ciancimino, E., Cannella, S. (2009). Modelling the bullwhip effect dampening practices in a limited capacity production network. *Lecture Notes in Business Information Processing,* 20, 475-486.

Crainic, T. G., Gobbato, L., Perboli, G., Rei, W. (2016). Logistics capacity planning: A stochastic bin packing formulation and a progressive hedging meta-heuristic. *European Journal of Operational Research*, 253(2), 404-417.

De Matta, R. (2017). Product costing in the strategic formation of a supply chain. *Annals of Operations Research*, 1-39. DOI: 10.1007/s10479-017-2463-x

De Souza, R., Zice, S., Chaoyang, L. (2000). Supply chain dynamics and optimization. *Integrated Manufacturing Systems*, 11(5), 348-364.

Dejonckheere, J., Disney, S. M., Lambrecht, M. R., Towill, D. R. (2003). Measuring and avoiding the bullwhip effect: A control theoretic approach. *European Journal of Operational Research*, 147(3), 567-590.

Disney, S. M. Lambrecht, M.R. (2008). On replenishment rules, forecasting and the bullwhip effect in supply chains. Foundations and Trends in Technology, *Information and Operations Management*, 2(1), 1–80.

Disney, S. M., Towill, D. R. (2003). On the bullwhip and inventory variance produced by an ordering policy. *Omega*, 31(3), 157-167.

Dominguez, R., Cannella, S., Framinan, J.M. (2015a). On returns and network configuration in supply chain dynamics. *Transportation Research Part E: Logistics and Transportation Review*, 73, 152-167.

Dominguez, R., Cannella, S., Framinan, J.M. (2015b). The impact of the supply chain structure on bullwhip effect. *Applied Mathematical Modelling*, 39 (23-24), 7309-7325.

Evans, G. N., Naim, M. M. (1994). The dynamics of capacity constrained supply chains. Paper presented at the *International System Dynamics Conference*, Stirling, 28-35.

Evers, P.T., Wan, X. (2012). Systems analysis using simulation. *Journal of Business Logistics*, 33 (2), 8089.

Forrester, J. W. (1958). Industrial dynamics: a major breakthrough for decision makers. *Harvard Business Review*, 36(4), 37-66.

Fowler, J. W., Park, S., MacKulak, G. T., Shunk, D. L. (2001). Efficient cycle time-throughput curve generation using a fixed sample size procedure. *International Journal of Production Research*, 39(12), 2595-2613.

Framinan (2017). Capacity considerations in the bullwhip effect in supply chains: The effect on lead times. Paper presented at the *11th Conference on Stochastic Models of Manufacturing and Service Operations*. June 4-9, 2017, Acaya, Italy.

Fransoo, J. C., Lee, C. Y. (2013). The critical role of ocean container transport in global supply chain performance. *Production and Operations Management*, 22(2), 253-268.

Fromm, H. (1992). Some remarks on cycle time, variability, zero inventories, and costs in microelectronics manufacturing lines. *IBM Technical Report TR 28.167*, German Manufacturing Technology Center, Sindelfingen, Germany

Grubbström, R. W., Wang, Z. (2000). Introducing capacity limitations into multi-level, multi-stage production-inventory systems applying the input-output/Laplace transform approach. *International Journal of Production Research*, 38(17), 4227-4234.

Hamdouch, Y. (2011). Multi-period supply chain network equilibrium with capacity constraints and purchasing strategies. *Transportation Research Part C: Emerging Technologies*, 19(5), 803-820.

Hax, A. C., Candea, D. (1984). *Production and Inventory Management*, Prentice-Hall, Englewood, Cliffs, NJ.

Helo, P. T. (2000). Dynamic modelling of surge effect and capacity limitation in supply chains. *International Journal of Production Research*, 38(17), 4521-4533.

Hernandez, R. Vollmer, L. (1998), Effect of flexible capacities on logistic objectives. Paper presented at the 3rd Annual International Conference on Industrial Engineering Theories, Applications and Practice, December 28-31, 1998, Hong Kong.

Holweg, M. (2005). The three dimensions of responsiveness. *International Journal of Operations & Production Management*, 25(7), 603-622.

Holweg, M. and Disney, S.M., 2005. The evolving frontiers of the bullwhip problem. Paper presented at EUROMA 2005 conference. Budapest: Hungary, 707–716.

Hum, S. H., Parlar, M. (2014). Measurement and optimization of supply chain responsiveness. *IIE Transactions*, 46(1), 1-22.

Hussain, M., Khan, M., Sabir, H. (2016). Analysis of capacity constraints on the backlog bullwhip effect in the two-tier supply chain: a Taguchi approach. *International Journal of Logistics Research and Applications*, 19(1), 41-61.

Juntunen, J., Juga, J. (2009). Controlling the bullwhip with transport capacity constraints. *International Journal of Services and Standards*, 5(2), 160-171.

Kacar, N. B., Mönch, L., Uzsoy, R. (2016). Modeling Cycle Times in Production Planning Models for Wafer Fabrication. *IEEE Transactions on Semiconductor Manufacturing*, 29(2), 153-167.

Kelton, W.D., Sadowki, R.P., Sadowki, D.A. (2007). Simulation with Arena, fourth ed. McGraw-Hill, New York.

Kleijnen, J. P. C. (2008). Simulation experiments in practice: statistical design and regression analysis. *Journal of Simulation*, 2 (1), 19–27.

Leach SE, Fowler JW, Mackulak GT et al. (2005) Asymptotic variance-based sampling for simulating cycle time-throughput curves. *Working paper ASUIE-ORPS-2005–003*, Arizona State University.

Lee, H. L. (2010). Taming the bullwhip. *Journal of Supply Chain Management*, 46(1), 7-8.

Lee, H. L., Padmanabhan, V., Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, 43(4), 546-558.

Little, J. D. (1961). A proof for the queuing formula: L= λ W. *Operations Research*, 9(3), 383-387.

Mahapatra, S. K., Das, A., & Narasimhan, R. (2012). A contingent theory of supplier management initiatives: effects of competitive intensity and product life cycle. *Journal of Operations Management*, 30(5), 406-422.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2), 111-153.

Miragliotta, G. (2006). Layers and mechanisms: A new taxonomy for the bullwhip effect. *International Journal of Production Economics*, 104(2), 365-381.

Mönch, L., Fowler, J. W., Mason, S. (2013). Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis, and systems (Vol. 52). Springer Science Business Media.

Nemoto, K., Akcali, E., Uzsoy, R. M. (2000). Quantifying the benefits of cycle time reduction in semiconductor wafer fabrication. *IEEE Transactions on Electronics Packaging Manufacturing*, 23(1), 39-47.

Nepal, B., Murat, A., Chinnam, R. B. (2012). The bullwhip effect in capacitated supply chains with consideration for product life-cycle aspects. *International Journal of Production Economics*, 136(2), 318-331.

Orcun, S., Uzsoy, R., Kempf, K. G. (2009). An integrated production planning model with load-dependent lead-times and safety stocks. *Computers Chemical Engineering*, 33(12), 2159-2163.

Pahl, J., Voß, S., Woodruff, D. L. (2007). Production planning with load dependent lead times: an update of research. *Annals of Operations Research*, 153(1), 297-345.

Park, S., Fowler, J. W., Mackulak, G. T., Keats, J. B., Carlyle, W. M. (2002). D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve. *Operations Research*, 50(6), 981-990.

Ponte, B., Wang, X., de la Fuente, D., Disney, S. M. (2017). Exploring nonlinear supply chains: the dynamics of capacity constraints. *International Journal of Production Research*, 55(14), 4053-4067.

Reichhart, A., Framinan, J. M., Holweg, M. (2008). On the link between inventory and responsiveness in multi-product supply chains. *International Journal of Systems Science*, 39(7), 677-688.

Reichhart, A., Holweg, M. (2007). Creating the customer-responsive supply chain: a reconciliation of concepts. *International Journal of Operations & Production Management*, 27(11), 1144-1172.

Riddalls, C. E., Bennett, S. (2002). The stability of supply chains. *International Journal of Production Research*, 40(2), 459-475.

Riddalls, C. E., Bennett, S., Tipi, N. S. (2000). Modelling the dynamics of supply chains. *International Journal of Systems Science*, 31(8), 969-976.

Schisgall, O. (1981). *Eyes on tomorrow: The evolution of Procter & Gamble*. Ferguson Publishing Company, New York.

Shukla, V., Naim, M. M. (2017). Detecting disturbances in supply chains-the case of capacity constraints. *International Journal of Logistics Management*, 28(2), 398-416.

Slack, N. (1987). The flexibility of manufacturing systems. *International Journal of Operations & Production Management*, 7(4), 35-45.

Spiegler, V. L., Naim, M. M. (2014). The impact of freight transport capacity limitations on supply chain dynamics. *International Journal of Logistics Research and Applications*, 17(1), 64-88.

Sterman, J. D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, 35(3), 321-339.

Sterman, J. D. (2000). *Business dynamics: systems thinking and modeling for a complex world* (No. HD30. 2 S7835 2000).

Sterman, J. D. (2006). Operational and behavioral causes of supply chain instability. In Villegas, Felipe A., Carranza, Octavio A. (Eds.), *The bullwhip effect in supply chains: a review of methods, components and cases*. Palgrave Macmillan, London.

Sterman, J.D., Dogan, G. (2015). "I'm not hoarding, I'm just stocking up before the hoarders get here.": Behavioral causes of phantom ordering in supply chains. *Journal of Operations Management*, 39-40, 6-22.

Towill, D. R. (1996). Time compression and supply chain management-a guided tour. *Supply Chain Management: An International Journal*, 1(1), 15-27.

Udenio, M., Fransoo, J.C., Peels, R. 2015. Destocking, the bullwhip effect, and the credit crisis: Empirical modeling of supply chain dynamics. *International Journal of Production Economics*, 160, 34-46.

Upasani, A., Uzsoy, R. (2008). Incorporating manufacturing lead times in joint production-marketing models: a review and some future directions. *Annals of Operations Research*, 161(1), 171-188.

Veeger, C. P. L., Etman, L. F. P., Van Herk, J., Rooda, J. E. (2010). Generating cycle time-throughput curves using effective process time based aggregate modeling. *IEEE Transactions on Semiconductor Manufacturing*, 23(4), 517-526.

Vlachos, D., Tagaras, G. (2001). An inventory system with two supply modes and capacity constraints. *International Journal of Production Economics*, 72(1), 41-58.

Wikner, J. (2003). Continuous-time dynamic modelling of variable lead times. *International Journal of Production Research*, 41(12), 2787-2798.

Wilson, M. C. (2007). The impact of transportation disruptions on supply chain performance. *Transportation Research Part E: Logistics and Transportation Review*, 43(4), 295-320.

Wright, D., Yuan, X. (2008). Mitigating the bullwhip effect by ordering policies and forecasting methods. *International Journal of Production Economics*, 113(2), 587-597.

Xie, W., Jiang, Z., Zhao, Y., Hong, J. (2014). Capacity planning and allocation with multi-channel distribution. *International Journal of Production Economics*, 147, 108-116.

Yang, F., Ankenman, B., Nelson, B. L. (2007). Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Research Logistics*, 54(1), 78-93.

Yang, T., Wen, Y.-F., Wang, F.-F. (2011). Evaluation of robustness of supply chain information-sharing strategies using a hybrid Taguchi and multiple criteria decision-making method. *International Journal of Production Economics*, 134 (2), 458-466.