



Defining accurate delivery dates in make to order job-shops managed by workload control

Davide Mezzogori¹ · Giovanni Romagnoli¹ · Francesco Zammori¹

Accepted: 28 September 2020 / Published online: 23 October 2020
© The Author(s) 2020

Abstract

Workload control (WLC) is a lean oriented system that reduces queues and waiting times, by imposing a cap to the workload released to the shop floor. Unfortunately, WLC performance does not systematically outperform that of push operating systems, with undersaturated utilizations levels and optimized dispatching rules. To address this issue, many scientific works made use of complex job-release mechanisms and sophisticated dispatching rules, but this makes WLC too complicated for industrial applications. So, in this study, we propose a complementary approach. At first, to reduce queuing time variability, we introduce a simple WLC system; next we integrate it with a predictive tool that, based on the system state, can accurately forecast the total time needed to manufacture and deliver a job. Due to the non-linearity among dependent and independent variables, forecasts are made using a multi-layer-perceptron; yet, to have a comparison, the effectiveness of both linear and non-linear multi regression model has been tested too. Anyhow, if due dates are endogenous (i.e. set by the manufacturer), they can be directly bound to this internal estimate. Conversely, if they are exogenous (i.e. set by the customer), this approach may not be enough to minimize the percentage of tardy jobs. So, we also propose a negotiation scheme, which can be used to extend exogenous due dates considered too tight, with respect to the internal estimate. This is the main contribution of the paper, as it makes the forecasting approach truly useful in many industrial applications. To test our approach, we simulated a 6-machines job-shop controlled with WLC and equipped with the proposed forecasting system. Obtained performances, namely WIP levels, percentage of tardy jobs and negotiated due dates, were compared with those of a set classical benchmark, and demonstrated the robustness and the quality of our approach, which ensures minimal delays.

Keywords Delivery dates · Discrete event simulation · Job-shop · Workload control · Regression · Neural network

✉ Francesco Zammori
francesco.zammori@unipr.it

¹ Department of Engineering and Architecture, University of Parma, Viale G.P. Usberti, 181/A, 43124 Parma, Italy

1 Introduction

Nowadays, the successful application of lean manufacturing across industries of various sectors and with different characteristics, has reinforced the claim that lean is a universal production system that can bring a permanent competitive edge (Yadav et al. 2019). Especially in manufacturing and logistics, lean can help industrial practitioners to increase operational performance by developing a waste-free value stream, where jobs flow continuously from a value-added activity to the following one. The focus is on waste identification and removal and, in this regard, queues and inventories are considered the worst sources of wastes, as they increase cost and time, making delivery dates hardly predictable (Bertolini et al. 2013; Bhosale and Pawar 2019).

Using the words of Hopp and Spearman (2008), ‘Controlling Work-In-Process (WIP) and protecting throughput time from variance’ are two effective solutions to counteract the above-mentioned wastes. These basic concepts, however, are easily deployed in Make-To-Stock (MTS) production systems, where demand is often constant, or at least more predictable, and the production mix is relatively stable and not too differentiated. Indeed, several card-based systems, such as Dual-Kanban and CONWIP (acronym for constant Work in Process), have been implemented in MTS environments to provide simple visual solutions for production planning and control, and lead time reduction. Unfortunately, as several authors report (see for example Germs and Riezebos 2010; Harrod and Kanet 2013; Marangoni et al. 2013), the same approach does not provide comparable results in case of Make-To-Order (MTO) High-Variety-Low-Volumes (HVLV) manufacturers. According to Dörmer et al. (2013), customized MTO production is replacing the standardized MTS approach in important manufacturing sectors and, in those companies, even advanced techniques such as synchro-MRP or POLCA cannot guarantee satisfactory results, especially in case of complex and variable routings (Land 2009). More specifically, synchro-MRP systems work particularly well in general flow-shops where it is not possible to consistently reduce change-over times and level production using small batches, whereas POLCA can be also used in case of general job-shops, provided that routings are few and linear (Bertolini et al. 2017; Oosterman et al. 2000). As soon as the number of possible routings increases, the quantity of cards explodes, making these visual management solutions inapplicable.

In this context, an alternative viable solution is offered by workload control (WLC), a lean-oriented Production Planning and Control (PPC) system, which has been receiving a great deal of attention in the last decades (Thürer et al. 2011, 2017). Without the need of visual cards, WLC provides a hybrid push-pull PPC approach and it regulates the manufacturing system by means of timely and accurate shop floor data, which are generally collected with a Manufacturing Executing System or the like. As its name suggests, the amount of workload released to the (critical) machines of the system is continuously monitored, and new jobs are not released to the shop floor unless the workload meets some predefined criteria set by threshold values or norms. Specifically, if norms are properly fine-tuned,

the time spent in the shop floor by a job, namely the Shop Floor Throughput Time (SFTT), can be reduced and stabilized and, consequently, on-time deliveries can be boosted. In this regard, WLC may provide a clear competitive edge, especially if the capability to accept and to respect short due dates (DDs, we note that in our paper the terms due dates and delivery dates are used interchangeably) is a requirement of the market. This is typical for MTO manufacturers: too long DDs can cause the loss of customers' orders because their requirements cannot be met, whereas too short DDs can be hardly respected, and they can cause production planning difficulties, and penalties due to late order deliveries. For this reason, the main aim of several works dealing with WLC is to reduce tardiness, and especially the percentage of tardy jobs, without affecting the throughput rate of the system (Fernandes et al. 2016; Thüerer et al. 2017).

However, if WLC is regulated by simple operating rules, its performance, expressed in terms of on time deliveries, does not deviate much from that of a purely push operating system, with optimized dispatching rules. This happens mainly in case of job-shop systems with a rather low utilization rate (e.g. below 90%, see Bertolini et al. 2016a), typical of many Small and Medium sized Enterprises (SMEs). To counteract this drawback, alternative methodologies for workload aggregation and accounting over time have been proposed and tested (Akillioglu et al. 2016; Bergamaschi et al. 1997), sophisticated dispatching rules have been devised and the standard job-release mechanism has been improved by using, for instance, anti-starvation approaches (Fernandes et al. 2017). Although numerical simulations have often proved the quality of these approaches, we note that further complexity is added into a complicated PPC system to achieve an improvement in terms of on time deliveries. This makes WLC unattractive from an industrial perspective, as evidenced by the limited research that focused on practical implementations of WLC, especially in SMEs (Hendry et al. 2013; Stevenson et al. 2011).

Due to these issues, and starting from the observation that queuing time's variability can be substantially reduced even adopting simple jobs' releasing rules (Bertolini et al. 2016b), we tackle the problem in a complementary way. Specifically, rather than trying to further reduce the SFTT with complex rules, we keep WLC as simple as possible, and we exploit the stabilization that it produces to forecast the total time needed to process and to deliver new accepted jobs. Realistic DDs can then be bound to these estimations, and the probability of a job to be late can be reduced. This approach is not completely new and, indeed, some attempts of this kind have already been made in the literature, where a few simple rules for DDs definition can be found (Moreira and Alves 2009; Thüerer et al. 2019). Yet, as shown by Thüerer and Stevenson (2016a, b), these rules are not very performing, especially in case of HVLV job-shops, with complex routings.

To solve this criticality, the present study proposes to integrate WLC with a structured forecasting system that, starting from the analysis of the current system state, predicts jobs' throughput time as a function of the workload released to the (critical) machines. However, if delivery dates are exogenous, i.e. set by the customer, an accurate jobs' throughput time forecast could not be enough to minimize the percentage of tardy jobs. So, we also propose a negotiation scheme that can be used by

the manufacturer to extend delivery dates that are considered too tight, with respect to the internal estimates.

To test our approach, we simulated a 6-machines job-shop controlled with WLC, and we made throughput time forecasts using a multi-layer-perceptron, a standard deep learning approach that can be easily implemented in the industry. As we will explain later, the rationale behind this choice can be traced in the non-linearity among workloads (independent variables) and the expected total throughput times (dependent variable). The multi-layer-perceptron, in fact, is capable to handle non-linearity without requiring the definition, a priori, of a non-linear model, whose knowledge and/or determination could be a hurdle for a practical application of the method. Yet, for research purposes, and with the aim of providing a valid benchmark, we also tested the effectiveness of multiple linear and non-linear (i.e. polynomial) regression models.

Performances are measured in terms of percentage of tardy jobs and percentage of negotiated due dates; the results demonstrate the superiority of our approach, compared with a set of WLC configurations taken as benchmarks.

We note that a preliminary study that used this approach can be found in the conference paper presented in Berlin by Mezzogori et al. (2019), of which this work is an extension. However, the level of analysis of the present work is significantly higher than that of the previous conference paper and, most of all, many additional tests have been made and new operating conditions have been considered, so as to verify and make clear the operational aspects of our approach and the competitive advantages that it can provide to companies wishing to implement it. In detail, concerning technical elements related to WLC, if compared to the previous work by Mezzogori et al. (2019), in the present study we (i) analyze another job dispatching rule, namely the Operation Due Date, in addition to the already considered Earliest Due Date; (ii) add a different norm optimization method, beyond the previously used minimization of WIP levels, to find the norms that minimize the percentage of tardy jobs and (iii) checked whether or not a polynomial regression could be used in place of the multi-layer-perceptron.

Also, and perhaps more important, we particularly stress the importance of the negotiation strategy, which is extended and made more coherent in the present study with a real operating scenario. In this regard, we (i) introduce the possibility of negotiating due dates, considering three different scenarios: namely balanced market power between the manufacturer and the customer, manufacturer has more market power, and customer has more market power; (ii) provide both the Standard Push and the Standard WLC systems with two alternative negotiation methods, namely the blind and the selective negotiation, to assess whether the observed benefits are merely due to the negotiation method, or if they are also enhanced by a precise estimation of the Gross Throughput Time; (iii) assess performance even when the production capacity of the manufacturing system is almost saturated, a condition that is rather frequent in case of make to order job shops and (iv) introduce the reverse negotiation procedure, to provide the manufacturer with a system capable of reducing the due dates that are very late, if compared with the forecasted Gross Throughput Time. The last point might have an immediate practical return, because the capability to offer delivery dates that are shorter than the exogenous ones initially set by

the customers might be a key element to win an order, to build customer loyalty and/or to benefit of extra margins related to early deliveries.

To summarize, our study contributes to the existing literature in two ways: (i) it shows that even a standard and simple WLC can assure a competitive edge, if delivery dates are based on the estimates made by a robust forecasting system (ii) it presents a negotiation scheme that exploits the overall control system, ensuring minimal delays and making WLC effective and easy to use in the industry.

The remainder of the paper is organized as follows. Section 2 provides an introduction to WLC research, and the main features of the forecasting procedure are detailed in Sect. 3. Section 4 presents the simulated system, the experimental campaign and the obtained results. In Sect. 5, we provide deeper details on the due date negotiation process, together with its application in a real operating scenario. Lastly, concluding remarks and possible directions for future works are drawn in Sect. 6.

2 An overview on WLC research

WLC is a PPC system proposed for the first time in 1981 to solve the so-called ‘lead time syndrome’ (Bertrand 1981), as its main goal is to ensure high throughput rates, as well as short and stable SFTT (Bertrand and Van Ooijen 2002; Henrich et al. 2007). As Stevenson et al. note (2005), this goal is particularly relevant in case of HVLV manufactures that often produce according to MTO logic and that must provide a high level of mix-flexibility, without compromising short and stable SFTT. Indeed, those companies often maintain ‘batches and queues’ system, because mix-flexibility generally requires a job-shop system with WIP accumulating between machines, with the pressing need to balance production and to stabilize material flows as much as possible. To this aim, WLC decouples the shop floor from the flow of incoming jobs, and new jobs are released into production only if some specific workload norms are respected. Norms are generally expressed in terms of expected working hours, on which an upper bound is set; yet other options are possible, such as the use of a lower bound or even of both an upper and lower bound (Bergamaschi et al. 1997). If norms are not met, jobs are not released and they remain in a Pre-Shop-Pool (PSP) of pending orders, at least until the next consideration cycle takes place. We note, in this regard, that the time spent in the PSP (namely the PSP_{time}) is not negligible and it always accompanies the SFTT, thus creating the so-called Gross Throughput Time, $GTT = (PSP_{time} + SFTT)$.

More precisely, as shown by Fig. 1, WLC makes use of three levels of control, that are: (i) Job Entry, (ii) Job Release and (iii) Job Dispatching (Bertolini et al. 2016a).

2.1 Job entry

At this stage, two control decisions are performed, namely (i) order acceptance/rejection and (ii) Due Date definition (Moreira and Alves 2009). Both decisions can undergo a negotiation process between the manufacturer operating the WLC

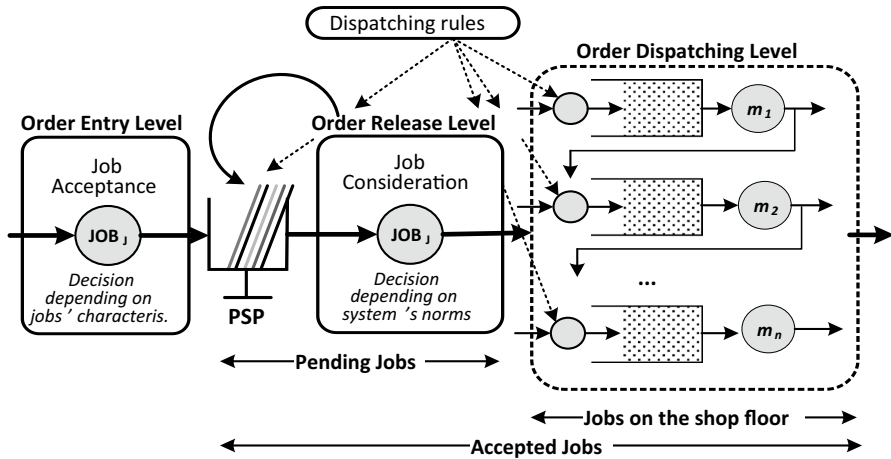


Fig. 1 WLC three levels of control

system and its customers. More precisely, customers' requests are assessed in terms of technical and economic feasibility and a DD is also defined, according to which the final acceptance decision is made. If a job is accepted, it is inserted in the PSP, waiting to be released to the shop floor.

The school of research where the arrival process of incoming orders is controlled by the decision on acceptance/rejection started with studies on queueing theory already 50 years ago (Miller 1969; Scott 1970), but this research branch is relatively small, and the control approach it proposes is extreme (Nandi and Rogers 2003). More recently, the order acceptance decision and the due date definition have been treated in parallel, with the negotiation option; if orders are accepted, in fact, the due date assignment decision is made immediately after the acceptance decision (Moreira and Alves 2006a). Common approaches that deal with the order acceptance and due date assignment problem are the total acceptance (i.e. no rejection decision is possible), the acceptance based on the present and future workload (Nandi 2000), and the due date negotiation method (Moreira and Alves 2006b).

It is worth noting that Thürer et al. (2019) have recently proposed an equation to generate feasible due dates via forward scheduling. This equation is composed of two elements, or 'lead time allowances', which account for the time spent by a job in the PSP and in the shop floor, respectively. In addition to these elements, that are dynamically updated depending on the job's routing and the current system state, a third and constant allowance is also considered, to compensate for possible deviations between the estimated lead time and the actual delivery time. However, as clearly stated by Thürer et al. (2019), the equation they propose is only meant to generate, via simulation, realistic endogenous due dates. It is not used in any way to make forecasts that will be used as the starting point of a negotiation process, as, instead, we will suggest in the present work.

2.2 Job release

Orders in the PSP are sorted using a dispatching rule and they are individually considered for possible release, either continuously or periodically, i.e. at predefined intervals of time. Anyhow, the decision depends both on the workload of the considered job and on the system’s norms, as explained below.

Let $w_{j,m}$ be the workload, generally expressed in working hours, of job j on machine m . When a job j^* is considered for possible release at time t , its workload is used to update the overall workload seen by the system W_t and/or the specific workload $W_{m,t}$ released to one or more (critical) machines. This is done according to Eqs. (1) and (2):

$$W_{m,t^+} = W_{m,t} + w_{j^*,m} \tag{1}$$

$$W_{t^+} = W_t + \sum_m w_{j^*,m} \tag{2}$$

where t^+ is the time immediately after job j^* has been considered for release and $w_{j^*,m} = 0$ if machine m does not belong to the routing of j^* .

Concerning $W_{m,t}$ and W_t , these quantities are computed as in Eqs. (3) and (4):

$$W_{m,t} = \sum_{j \in J_{(S,m)}} w_{j,m} \tag{3}$$

$$W_t = \sum_m \sum_{j \in J_{(S,m)}} w_{j,m} = \sum_m W_m \tag{4}$$

where $J_{(S,m)}$ is the set of the jobs that are already in the shop floor and that still have to visit machine m .

It is worth mentioning that Eqs. (1) and (2) are used any time a job is considered for release; instead, Eqs. (3) and (4) are used to update the system state every time a machine has processed a job and thus the job joins the queue of the next machine in its routing.

After the update, W_{t^+} and/or W_{m,t^+} are compared to the system’s norms and if all norms are respected, job j^* is released to the shop floor, otherwise it remains pending in the PSP. Whether job j^* has been released or not, the next one in the PSP is considered, and the process is iterated until the whole list of job has been completed.

Relatively to the comparison of the updated workload and the system’s norms, we note that norms are generally considered as a fixed upper bound, but other options are also common, such as lower bound or upper and lower bound. Also, according to Bergamaschi et al. (1997), the use of a single norm to control the workload of the system, i.e. W_t , is labelled *total shop load*, whereas the use of multiple norms to limit the workload $W_{m,t}$ of single machines is called *bottle-neck load* or *load by each machine*, if control is limited to the sole bottleneck or extended to all machines, respectively.

We also note that an important decision concerns the way in which workloads $w_{j,m}$ are quantified. In general, given a machine m , its workload can be separated into a direct and an indirect component. The first part is due to those jobs that are currently queuing in front of machine m , while the second part is that of the jobs that shall be processed by machine m , but are presently queuing at another machine upstream of m . If job j is waiting in the queue before machine m , its workload contribution $w_{j,m}$ is equal to the sum of its processing time and set up time. On the contrary, if job j is queuing upstream of m , $w_{j,m}$ should be reduced to consider the smaller urgency of the workload contribution of the job j on machine m . Specifically, only a portion of the workload contribution of job j should be added to the workload of machine m ($\tilde{w}_{j,m} \leq w_{j,m}$): the more m is downstream, the fewer load should be attributed to it (Wiendhal 1995).

A first way to dynamically rescale $w_{j,m}$, namely the Load Oriented Order Release (LOOR) approach, was firstly proposed by Bechte (1988, 1994), who suggested using a depreciation factor based on historical data. After the seminal works of Bechte, other probabilistic approaches have been proposed in technical literature. For example, Land and Gaalman (1998) introduced the Superfluous Load avoidance Release procedure, Cigolini and Portioli-Staudacher (2002) suggested the workload balancing method. Although these methods have reported interesting results in simulative environments, they have been mostly neglected in recent years and have seldom found their way into practice (Thürer et al. 2011). Probably, as noted by Stevenson (2006), these methods are over-sophisticated and, for this reason, they have been misused through lack of understanding or neglected over time.

Owing to this, the so-called *aggregate approach*, in which the direct and the indirect workload are simply added together, is generally applied (Thürer et al. 2011; Thürer and Stevenson 2016a). According to this simplified approach, $w_{j,m}$ is a fixed valued (equal to the processing and set up time of job j on machine m) that is never rescaled, not even if job j is queuing at a machine upstream of m . As an alternative, also the *corrected aggregate approach*, proposed by Land and Gaalman (1996), is frequently applied. In this case $w_{j,m}$ is rescaled in a simple way that does not require statistical data. Specifically, when job j is released, $w_{j,m}$ is rescaled as in Eq. (5), using as scaling factor the position $n_{j,m}$ of machine m in the routing of job j .

$$\tilde{w}_{j,m} = \frac{w_{j,m}}{n_{j,m}} \quad (5)$$

As proved by Oosterman et al. (2000), the corrected aggregate approach performs arguably better than the standard one, especially if a dominant flow exists.

2.3 Job dispatching

When jobs are released to the shop floor, they are moved from a machine to the following one as in a standard ‘batch and queue’ system. Whenever a queue is encountered, the job is added to the list of job competing for the same machine and the system must decide the order in which jobs of the list should be sorted for future processing. To this aim, a huge set of different dispatching rules can be used (van

Ooijen and Bertrand 2001): a comprehensive overview of available dispatching rules clearly falls outside the aim of this paper, and we refer the interested reader to the work of Chiang and Fu (2007).

3 Problem description and proposed approach

The capability to set and respect short DDs in response to customers' enquiries is a key success factor for MTO manufacturers. Generally, for a purely push job-shop system, the variability of the SFTT is very high, and the definition of accurate DDs is challenging, and it often bewilders production managers. This task could be simplified with the introduction of WLC, as this PPC system dramatically reduces work in process and stabilizes queuing times. However, as discussed in Sect. 2.1, WLC does not include a standard framework for DDs setting: even if the DD definition problem has been mentioned by many studies, a robust solution has not yet been presented, and only some simple and generic rules can be found in the literature. A first consistent attempt was made by Mezzogori et al. (2019) who propose integrating WLC with an effective forecasting system, based on statistic or machine learning techniques. In the present study, we continue and complete that preliminary study, by implementing and assessing the performances of a forecasting system based on a multi-layer-perceptron.

Specifically, any time a job enters the PSP, the current workload of the job-shop is observed and, based on this information, the forecasting system predicts the expected Gross Throughput Time (GTT), defined as the sum of the SFTT and of the waiting time in the PSP. Next, in case of endogenous DDs this prediction is immediately used to promise a feasible DD; otherwise, if DDs are exogenous, the prediction is used as a starting point of a negotiation process with the customer.

Therefore, to develop and operate the overall system, the following steps are needed:

- *WLC setting* To facilitate prediction, the variability of the GTT should be contained as much as possible. To this aim a WLC system should be deployed and properly configured. The main decisions here concern the selection of (i) a job release strategy, (ii) a workload quantification procedure and (iii) a proper dispatching rule.
- *Fine tuning of the norms* Both the number (i.e. single or multiple) and the type (i.e. lower or upper bound) of norms must be defined. Next, their value should be fine-tuned to optimize performances, without affecting the throughput rate of the job shop.
- *Development and fitting of the forecasting model* Proper regression variables must be selected, and the forecast model must be developed. Next, operating data must be collected to fit and validate the model.
- *Definition of a due dates generation and negotiation scheme* Lastly, to operate the system, a proper negotiation scheme must be defined and coupled with the forecasting system.

The first three points are explained in the following subsections. Conversely, possible due dates generation and negotiation schemes are introduced in Sect. 4.4, where numerical examples are also provided.

3.1 WLC setting

It is not possible to define a general rule for a proper configuration of the WLC system, as this decision heavily depends on the job-shop under analysis. Yet, in favor of an easy management of the system, we suggest using (cf. Bergamaschi et al. 1997):

- load limited order release mechanism, with discrete timing convention;
- load by each machine, with upper bound norms;
- workload calculated at each machine with the corrected aggregate approach, with passive capacity planning and limited schedule visibility;
- jobs sorted in the PSP and in machines' queues, either using the Earliest Due Date (EDD) or the Operation Due Date (ODD) dispatching rule.

Indeed, these choices are quite common in WLC literature, and they have proven to be very effective in reducing the GTT and, consequently, the percentage of tardy jobs. Also, although many papers simulated and compared alternative dispatching rules to optimize certain parameters in different production systems, we considered this issue of lesser importance. Quite often, in fact, simple dispatching rules are enough to generate interesting results, if coupled with optimized release and due date definition tools. Hence, with the aim of maximizing performances related to on-time delivery, we suggest using the EDD or the ODD rule, as discussed in Moreira and Alves, (2009) and in Fernandes et al. (2017). Briefly, we recall that the EDD is a priority rule that sequences jobs in a queue according to their due dates i.e. jobs with closer due date get higher priority. The ODD is also a time-based rule that considers the urgency of a job; in this case, however, the urgency is dynamically corrected as jobs proceed in the shop floor. This is clearly shown by Eq. (6):

$$ODD_{i,j} = DD_j - (M_j - i) \cdot c \quad \text{with } i = 1, \dots, M_j \quad (6)$$

where M_j is the total number of machines of the routing of job j , i is the current position of job j (on its routing) and c is a constant allowance factor, generally taken in the range [2–5].

3.2 Fine tuning of the norms

WLC configuration ends with the definition of a proper upper bound level, say N_m , of the norms regulating each machine m . This decision is known to be crucial, as it has the greatest impact on performance (Thurer et al. 2011; Thürer et al. 2014). To clarify this issue, we recall that, as explained in Sect. 2, as soon as job j is released in the job-shop, its workload contribution $w_{j,m}$ is added to the workload W_m of machine m , even if m is not the first machine on the routing of job j . Conversely, $w_{j,m}$ is subtracted from W_m only when job j leaves machine m and moves to the next machine

or exits the system. Basically, the workload contribution of a job accounts both for the direct and indirect load, as it immediately contributes to the workload of all downstream machines. Due to this issue, a specific norm N_m should be defined for each machine m and, consequently, finding the best combination of the norms would be extremely hard. This reasoning is certainly correct if the aggregate account of workload over time is used. Fortunately, as demonstrated by Thurer et al. (2011) and lately by Fernandes, Land, and Carmo-Silva (2014), who investigated the norms' optimization problem, when the workload is converted (i.e. rescaled) using the corrected aggregate approach, 'it dynamically adjusts itself to the current situation on the shop floor at any moment in time'. For this reason, there is no need to search an optimal combination of norms, but it is enough to use a same common norm N^* for each machine, i.e. $N_1 = N_2 = \dots = N_m = \dots = N^*$. This is another element that advocates for the use of the corrected aggregate approach.

Owing to this issue, N^* can be easily found with a straightforward exhaustive procedure, as detailed in Bertolini et al. (2016a). Specifically, the procedure starts from an initial and very high value of the norm N , which should assure that, as in a push system, jobs are never stopped and immediately released upon acceptance. Next, N is iteratively reduced stepwise down, using a constant and small step δ , until the average WIP or the percentage of tardy jobs (but other performance criteria could be used too), continues to decrease and/or the maximal throughput rate ρ remains almost unaltered. The last constraint is needed to prevent the system from losing part of its productive capacity, due to an over-restrictive jobs release phase. For the sake of clarity, a pseudo code is shown in Table 1.

Please note that the maximum workload W_m , registered when the job-shop operates in a push way, is used as the initial (very high) value of the common norm. This condition assures that, at the beginning of the iterating procedure, the WLC behaves exactly as a push system. Also note that the constraint on the throughput rate ρ is assessed statistically, performing a t test, at level $(1 - \alpha)$, at each (decreasing) level of the norm.

3.3 Development and fitting of the forecasting model

In general, any time a job is accepted and inserted in the PSP, the manufacturer should set a delivery date equal to the expected GTT plus a certain allowance factor, to account for possible deviations. This approach generates an internal, or endogenous DD, but there is no guarantee that a customer is willing to accept it. Anyhow, even if the manufacturer has not enough market power to impose the internal DD, this quantity still has a relevant value, as it can be the starting point to initiate a negotiation with the customer.

To estimate the GTT, we recall that this quantity is defined as the time between order acceptance and order delivery and, when WLC is used to regulate the system, it can be decomposed in the time spent by job j^* in the PSP and, next, in the shop floor. The latter one can be further partitioned in the processing, set up, and queuing time of job j^* , at each machine m of its routing R_{j^*} . This is clearly shown by Eq. (7):

Table 1 Pseudo code of the fine-tuning procedure, example on optimizing percentage of tardy jobs

```

# Input Values: M, R, k, α
# M is the number of machines
# R is the number of simulation runs, a possible value is 100
# k is the reducing factor for N, a possible value is 1%.
# (1 - α) is the confidence level, a possible value is 95%
# Preliminary Step – Push System
Set TJ* = 0
Set W = [ ], TR = [ ], TJ = [ ], TRWLC = [ ], TJWLC = [ ] # Empty Lists
Set N* = ∞ # WLC control is disabled
For i in Range(R):
    Perform a simulation run with norm equal to N*
    Compute the throughput rate ρ, and the Percentage of tardy jobs g
    TR.Add(ρ)
    TJ.Add(g)
    For m in Range(M):
        Compute Max(Wm) # The maximum workload released to m
        W.Add(Max(Wm))
Next i
Set TJ* = Mean(TJ)
Set N* = Max(W)
# Optimization Loop
Set δ = k · N*
While True:
    For i in Range(R):
        Perform a simulation run with norm equal to N*
        Compute the throughput rate ρ, and the Percentage of tardy jobs g
        TRWLC.Add(ρ)
        TJWLC.Add(g)
    Test E[TR] ≠ E[TRWLC] at level (1 - α) # t-test on the mean of Throughput rate
    If Test is positive: Break
    If Mean(TJWLC) < TJ*:
        Set TJ* = Mean(TJWLC)
        Set N* -= δ
        Set TRWLC = [ ], TJWLC = [ ]
    Else:
        Break
Return N*, TJ* #These are the optimal values

```

$$GTT_{j^*} = P_{j^*} + \left(Q_{j^*} + \sum_{m \in R_{j^*}} (s_{j^*,m} + p_{j^*,m}) \right) \tag{7}$$

where:

- P_{j^*} is the pending time in the PSP,
- Q_{j^*} is the queuing time in the shop floor,
- $s_{j^*,m}$ and $p_{j^*,m}$ are, respectively, the set-up and processing time,

- The terms in brackets $\left(Q_{j^*} + \sum_{m \in R_{j^*}} (s_{j^*,m} + p_{j^*,m}) \right)$ is the SFTT of job j^* .

In Sect. 2 we stressed the fact that, for the corrected aggregated approach, the workload $w_{j^*,m}$ coincides with the average processing and set up times of job j^* on machine m , divided by the position $n_{j,m}$ of machine m in the routing R_{j^*} . Owing to this fact, it is licit to substitute $(s_{j^*,m} + p_{j^*,m})$ with $(n \cdot w_{j^*,m})$ in Eq. (7). The approximation error introduced by this simplification is indeed very small, as the variability of both P_{j^*} and Q_{j^*} is expected to be much bigger than that of the total processing and set up time $\sum (s_{j^*,m} + p_{j^*,m})$. Consequently, since workloads $w_{j^*,m}$ are known, to obtain GTT_{j^*} we only need to estimate the waiting times P_{j^*} and Q_{j^*} . To this aim, we pose that whenever a job j^* is accepted and enters the PSP, both P_{j^*} and Q_{j^*} could be estimated as a function of the workloads of the other jobs pending in the PSP, and of the total workload released to the machines belonging to the routing of job j^* .

This is shown in Eq. (8), where only workload w_{j,m^*} , relatively to machine $m^* \in R_{j^*}$ are considered:

$$\begin{aligned}
 Y_{j^*} = & f_1 \left(\sum_{j \in J_p} \tilde{w}_{j,1^*}, \dots, \sum_{j \in J_p} \tilde{w}_{j,m^*}, \dots, \sum_{j \in J_p} \tilde{w}_{j,M^*} \right) \\
 & + f_2 \left(\sum_{j \in J(s,1^*)} \tilde{w}_{j,1^*}, \dots, \sum_{j \in J(s,m^*)} \tilde{w}_{j,m^*}, \dots, \sum_{j \in J(s,M^*)} \tilde{w}_{j,M^*} \right)
 \end{aligned}
 \tag{8}$$

where:

- Y_{j^*} is the total waiting time to be estimated,
- $\{1^*, \dots, m^*, \dots, M^*\}$ is the set of machines belonging to R_{j^*} ,
- J_p is the set of the jobs pending in the PSP,
- $J_{(s,m^*)}$ is the set of jobs in the job-shop that have not yet been processed by machine m^* ,
- $\left(\sum_{j \in J(s,m^*)} \tilde{w}_{j,m^*} + \sum_{j \in J_p} \tilde{w}_{j,m^*} \right) = \left(\sum_{j \in J_p} \tilde{w}_{j,m^*} + W_{m^*} \right)$ is the cumulated workload of machine m^* , due to the jobs in the PSP and to the ones in the job-shop.

Different approaches could be used to fit Eq. (8) and, among the different alternatives, we suggest using a Multi-Layer-Perceptron (MLP), a feed forward fully connected neural network, with at least one hidden layer. At present, MLP is a well-established tool, with many scientific and industrial applications; hence, we believe that it cannot be considered as a hurdle for the future implementation of our method in industrial practice. Apart from that, the rationale behind this choice can be traced in the famous Little’s Law (Little 1961), which indicates a non-linear relation between the GTT and the WIP. According to the Universal Approximator Theorem (see for example Tikik et al. 2003), in fact, the multi-layer-perceptron is known to better approximate and exploit non-linear relationships between target and input variables. Whenever non-linearity is present, the

neural network can automatically trace it during the training, without any help from human experts. In other words, unlike other forecasting models such as a polynomial regression, there is no need to preliminary guess the form of non-linearity among the variables. This is a clear benefit for practitioners, with the modest down-side of a slightly more complex architecture.

Nonetheless, we will also consider a Multiple Linear Regression (MLR); there are, indeed, at least two reasons to do so. First, the MLR provides an immediate benchmark solution; second, it allows a straightforward assessment of the significance level of the selected regressor variables. Using a neural network, a similar analysis is possible (see for example Montavon et al. 2018), but it is a rather convoluted task, which is not yet fully accepted by academics. A further discussion on this topic is postponed in Sect. 4, where a numerical example is provided.

Anyhow, in both cases, to fit the models, the following quantities must be collected (or virtually generated), for each accepted job j^* :

- the time t_{Acc,j^*} when the job j^* is accepted and inserted in the PSP,
- its workload contribution \tilde{w}_{j^*,m^*} , for all the machines m^* of its routing R_{j^*} ,
- the cumulated workloads in the PSP and in the shop floor (i.e. $\sum_{j \in J_p} \tilde{w}_{j,m^*}$ and W_{m^*}) observed immediately before job j^* is accepted at time t_{Acc,j^*} ,
- the time t_{End,j^*} when job j^* is completed and leaves the job-shop.

Specifically, $\sum_{j \in J_p} \tilde{w}_{j,m^*}$ and W_{m^*} are used as regressors or independent variables, whereas the dependent variable Y_{j^*} is quantified as $(t_{End,j^*} - t_{Acc,j^*}) - \sum_{m \in R_{j^*}} (n_{j^*,m} \cdot w_{j^*,m})$.

We conclude this section observing that, once the forecasting system is in place, real operating data, properly collected and stored, could be used to dynamically update and to refine the forecasting models.

4 Models implementation and assessment

4.1 The WLC simulated job-shop

To test the forecasting model and the related negotiation scheme, we reproduced a High-Variability Low Volumes (HVLV) six-machines job-shop. To this aim we used Simpy®, an open source discrete event simulation package, developed in Python 3.7©. The six-machines job-shop that we considered is typically used as benchmark, in most of the scientific works dealing with WLC (Bertolini et al. 2016b). Specifically:

- machines are equally loaded, with an average utilization rate or 90% and they have constant capacity and full availability i.e. time losses due to failures and reparations are not considered;

- the job shop is ‘pure and randomly routed’, in that routings may include a random number of machines (from one to six), visited in random order; yet machines cannot be visited more than once;
- processing time $p_{j,m}$ (of each job j on machine m) follows a 2-Erlang distribution truncated at 4, with a mean \bar{p} of 1.0-time units. This choice is typical to reproduce the processing time variability of HVLV job-shops;
- set-up times are not sequence dependent, but set-up must be performed any time a machine processes a new job. Hence, set-up times are not explicitly defined, but they are included in the processing times $p_{j,m}$ generated from the 2-Erlang distribution, as described above;
- jobs are generated according to a Poisson process, with exponential distributed inter-arrival times, with arrival rate λ of 1.54 units of time. This value, together with the average processing time $\bar{p}=1.0$, assures the desired 90% utilization rate of each machine;
- an exogenous DD, representing the DD proposed or requested by the customer, is assigned to each generated job. Its value is randomly generated in the interval [28, 56], following the same logic proposed by ample WLC research (see for example Land 2006; Thürer et al. 2019).

Concerning the WLC used to regulate the job-shop, the same settings suggested in Sects. 3.1 and 3.2 were used, as summarized in Table 2. We just note that, in case of ODD rule, the allowance factor c reported in Eq. (6) was set to 4 units of time. This value was determined with a series of preliminary simulation runs, as the one that minimizes the percentage of tardy jobs, when the job-shop operates in a purely push way (i.e. jobs are immediately released upon arrival). All these settings ensure a full comparability of the simulation model with most of the ones proposed in the literature (see for example Land 2006; Thürer and Stevenson 2016b).

Table 2 Job-shop and WLC parameters

Number of machines	6 machines
Routing	Random length (1–6) with random sequence, no re-entrant loops
Operation time	2-Erlang distribution truncated at 4 with mean of 1-time unit
Set-up time	Included in the operation time
Machines’ utilization	90% for all machines
Machines’ availability	100% for all machines
Job’s arrival rate	Exponentially distributed, 1 job every 1.54-time units.
WLC jobs acceptance approach	All jobs are assumed to be feasible, and the ‘total acceptance’ criterion is used
WLC jobs consideration approach	Discrete timing convention, with jobs in the PSP evaluated every 24 units of time
WLC jobs release approach	Load by each machine, with corrected aggregated load and upper bound norms
Dispatching rules	EDD and ODD
Exogenous due dates	Uniformly distributed in the interval [28, 56]

4.2 Norms level

The use of six equal norms $N_1 = N_2 = \dots = N_6 = N^*$ is a sound solution for a perfectly balanced six-machines job-shop. This choice is also motivated by the use of the corrected aggregate load, as explained in Sect. 3. In the present study we will find the value N^* that minimizes either the percentage of tardy jobs or the WIP level. The first optimization criterion is commonly used in WLC literature, as a very high number of on time deliveries is often considered as an essential market requirement for MTO manufacturers. WIP minimization is, instead, one of the main objectives of lean manufacturing, as this condition minimizes queues and SFTT, with clear operational benefits, such as: lower holding costs, smaller occupied spaces and greater tidiness on the shop floor, and easier traceability. Nonetheless, this objective is rarely considered in WLC literature, essentially because WIP minimization makes the jobs' release phase very restrictive as the system becomes very constrained, leading to greater waiting time in the PSP. Consequently, the GTT becomes comparable to, if not even longer than, the SFTT of a purely push operating job-shop, and most of the improvements in terms of on time deliveries get lost. Yet, we investigated also this optimization criterion because, due to the introduction of a forecasting system and of a negotiation phase for DDs definition, we believe that on time deliveries could be increased even when WIP is minimized, thus obtaining a double operating advantage.

Since the selected dispatching rules has a direct and non-negligible impact on the optimal value of the norms, to find N^* the stepwise procedure described in Sect. 3.2 was repeated twice, using the EDD and the ODD dispatching rule, respectively. In both cases, we used simulation trials of 100 runs, each of 3650 units of time. A warm-up period of 1200 units of time, enough to reach the steady state of the system, was also included in each simulation run. These simulation parameters are consistent with those generally used in case of job-shop simulations and provide stable results in an adequate amount of time.

The 95% confidence interval of the throughput rate ρ (of the push operating job shop) was found at 1.525 ± 0.037 [jobs/unit of times] and the initial high level of the norm was set at 20-time units. Next, using a fixed decrement $\delta=1$, the best operating points, for each couple of dispatching policy and optimization performance, were easily found. These points are graphically shown on the curves of Fig. 2, which display the percentage of tardy jobs, for both dispatching rules, at each level of the norms over the entire evaluated range [3, 20]. In the same figure, a vertical dashed line is used to mark the threshold limit of the throughput rate i.e. points on its right respect the constraint on the throughput rate, while those on the left violate it. Clearly, the points highlighted in blue, located at the minimum of the curves, are the ones minimizing the percentage of tardy jobs. Instead, the ones highlighted in green, located at the immediate right of the vertical dashed line, correspond to the minimum level of WIP needed to sustain production, at its maximum throughput rate. This is a direct consequence of the famous Little's Law which states that the highest is the SFTT the highest is the WIP and vice versa. So, in Fig. 2, WIP decreases moving from right to left and the points located to the left of the dashed line reduce WIP at the expense of the throughput rate, which drops below its target value.

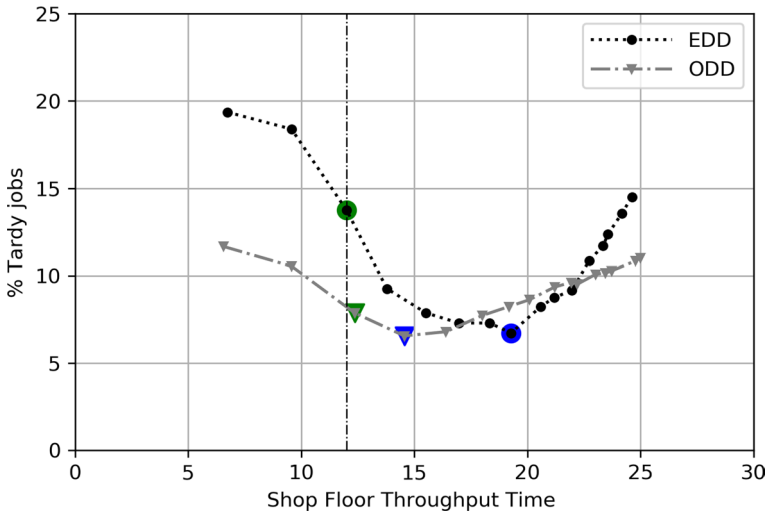


Fig. 2 Trend of % Tardy Jobs as a function of SFTT

For the sake of completeness, overall results are summarized in Table 3.

4.3 Forecasting models

As already anticipated, we predict that a non-linear relationship could exist between the GTT and WIP. Therefore, the main model we investigated is the MLP, whereas the MLR is only considered as a benchmark to measure the predictive performance of a simpler model which does not consider non-linear relationships between variables.

4.3.1 Multi-layer perceptron and multiple regression

The forecasting models were created using Scikit-learn© and Keras© with TensorFlow© as backend, two famous libraries for machine and deep learning (Géron 2019). Relatively to the six-machines job-shop, the relationship between the waiting

Table 3 Norms optimization results

Dispatching rule	System type	Norm level N*	Thr. rate ρ	WIP	% Tardy jobs	SFTT	GTT
EDD	PUSH	∞	1.525	128.84	13.94%	25.74	25.74
	WLC	5 (opt. WIP)	1.526	47.41	13.74%	12.01	31.18
		10 (opt. TARDY)	1.529	85.41	6.72%	19.27	25.00
ODD	PUSH	∞	1.528	96.83	10.61%	25.81	25.81
	WLC	5 (opt. WIP)	1.526	44.45	7.87%	12.36	26.11
		6 (opt. TARDY)	1.528	51.76	6.56%	14.56	23.08

time and the workloads formalized in Eq. (8), can be estimated by fitting the linear model as in Eq. (9):

$$Y_{j^*} = C + \left(\alpha_1 \sum_{j \in J_p} w_{j,1} + \dots + \alpha_6 \sum_{j \in J_p} w_{j,6} \right) + \left(\beta_1 \sum_{j \in J_{(s,1)}} w_{j,1} + \dots + \beta_6 \sum_{j \in J_{(s,6)}} w_{j,6} \right) \quad (9)$$

In Eq. (9), C is the intercept, whereas alpha and beta are 12 regression coefficients to be estimated. More precisely, alpha and beta coefficients refer to the workloads of the six machines due to the jobs in the PSP and in the shop floor, respectively.

Instead, to estimate Eq. (8) with a neural network, an MLP was used. More precisely, after performing an optimisation process, the following topology was obtained:

- input layer made of 12 neurons, one for each observed workload,
- single output neuron, returning the total expected waiting time,
- three fully connected hidden layers, each one made of 128 neurons plus a bias.

Also, neurons are activated using the Relu function and batch normalization is used, after each hidden layer.

4.3.2 Models fitting

To fit the models, we performed a total of 5000 simulation runs, each one with a duration of 4850 time-units, of which 1200 used as warm up. By doing so, we generated 22,430,954 observations: 80% of them were used as training set, the remaining 20% as test set.

Fitting the MLR we got an $R^2 = 0.4$, with α and β coefficients as shown in Table 4.

These values refer to the case of WLC with EDD dispatching rule and norms optimized to minimize WIP; yet, we found similar results also in other configurations. Although R^2 is rather low, we did not try to improve it using a higher degree polynomial and/or considering possible interaction effects. The aim of the MLR was, indeed, to provide a benchmark forecast system; instead, the search for possible non-linear effects was left to the MLP.

Nonetheless, it is important to note that all coefficients are significant, with a very low P -value. This certifies the correctness of the regressor variables that we choose, and it justifies their use in the neural network too. Also note that the alpha values are very similar, as are the beta ones. This result is totally coherent with the simulated scenario, as the six machines are identical with the same utilization level, and jobs' routings are totally random. Even the fact that alpha coefficients are higher than the beta ones is logical. Indeed, since norms were optimized to minimize WIP, the jobs

Table 4 Regression coefficients and statistical evidence (values, confidence intervals and *P* values)

Coefficient	Value	Confidence interval	<i>P</i> value
<i>C</i>	8.9908	[8.910–9.001]	< 1.0 E–3
Coefficients of PSP Workloads			
α_1	0.1568	[0.156–0.158]	< 1.0 E–3
α_2	0.1556	[0.155–0.156]	< 1.0 E–3
α_3	0.1538	[0.153–0.155]	< 1.0 E–3
α_4	0.1551	[0.154–0.156]	< 1.0 E–3
α_5	0.1560	[0.155–0.157]	< 1.0 E–3
α_6	0.1590	[0.158–0.160]	< 1.0 E–3
Coefficients of shop floor workloads			
β_1	0.1459	[0.145–0.147]	< 1.0 E–3
β_2	0.1435	[0.143–0.144]	< 1.0 E–3
β_3	0.1450	[0.147–0.153]	< 1.0 E–3
β_4	0.1446	[0.144–0.145]	< 1.0 E–3
β_5	0.1439	[0.143–0.145]	< 1.0 E–3
β_6	1.1447	[0.144–0.152]	< 1.0 E–3

release phase is very binding and the waiting time in the PSP has an impact (on the GTT) greater than that of the time spent in the shop floor.

Concerning the MLP, the training was made using the back propagation algorithm, based on the Adam optimizer (Kingma and Ba 2014). The latter is an algorithm for first-order gradient-based optimization, which takes advantage of adaptive estimates of lower-order moments. Specifically, the training was based on 500 epochs, with early stopping. A patience of 50, a validation split of 5% and a batch size of 2048 were used too. Lastly, all hyperparameters were found through cross-validation as suggested by Hastie et al. (2009).

4.3.3 Models validation

Using the MLR to perform forecasts on data of the test-set, a Root Mean Square Error (RMSE) of 13.43 time units was found. Conversely, as we expected by using the MLP, the RMSE dropped to 12.01 units of time, confirming the non-linear dependence between response and explanatory variables.

For the sake of clarity, Fig. 3 also shows the density estimation and the comparison between true and predicted values of the GTT, obtained either with the linear regression (MLR) or neural network (MLP). As already measured in terms of RMSE, density comparison shows a slight difference between neural network and linear regression, where the latter poses most mass in correspondence of the average value of the true distribution, and it slightly underestimate the mass of the right tail.

Lastly, to further verify the accuracy of our model, we tried to fit the experimental data using a polynomial model. It is worth noting that we used this model only for validation purpose, but we do not suggest it as a valid alternative of the MLP. Indeed, such predictive model would require, from an operational and implementational

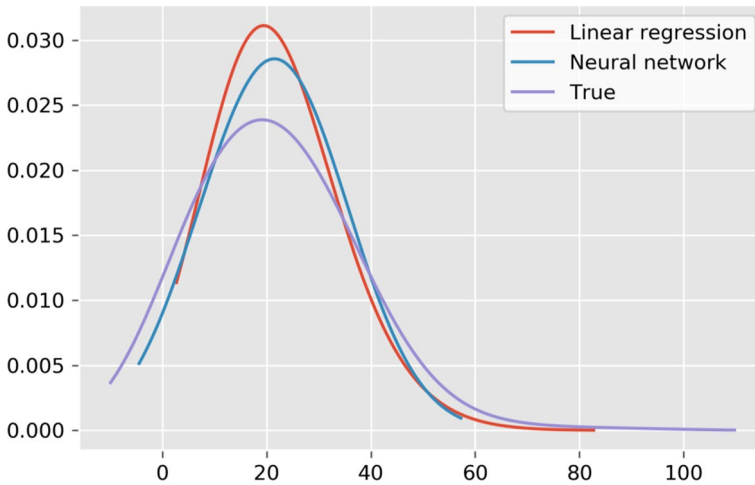


Fig. 3 Density estimation and comparison between predictions and true values

point of view, a highly time-consuming activity of feature selection (i.e. backward/forward feature selection), based on expert intuition of the model order. Due to this issue, we limited the analysis to a full factorial quadratic model: we have considered all the terms of both degree 1 and 2, as well as all the interaction factors, for a total of 90 independent variables. The obtained RMSE equals 13.20 units of time, with an $R^2 = 0.44$. Also, all coefficients of both the linear and quadratic terms were statistically significant, as well as the interaction terms between the workload of a machine m in the PSP and in the shop floor. Hence, non-linearity is statistically confirmed, but the increase of R^2 is limited, especially considering the much greater complexity of the regression model. Also, the obtained RMSE is worse than that of the MLP, a fact that confirms the superiority of the neural network approach.

4.4 Definition of a due dates generation and negotiation scheme

To assess the quality of the forecasting models we repeated the simulations of the WLC system in each of the four optimal configurations, as defined in Table 3. At this point, however, we added the forecasting system, to get a robust estimation of the GTT of the incoming jobs, and we also introduced a negotiation process between the manufacturer and the customer. Specifically, to reproduce a plausible operating scenario, any time the exogenous due date (DD_{ex}) requested by the customer is too tight, relatively to the estimated GTT, a negotiation process starts, and the manufacturer tries to extend DD_{ex} as much as possible. In detail, the negotiation process works in the following way:

- any time a job j is accepted and inserted in the PSP, an exogenous due date is created. It corresponds to the delivery date requested by the customer and its value is generated as a random number uniformly distributed on the interval [28, 56], as explained in Sect. 4.1;

- at the same time, the workload of the system is observed, and the gross throughput time of the incoming job j is estimated as: $GTT_j = \left(Y_j + \sum_{m \in R_{j^*}} (w_{j,m}) \right)$, where the total waiting time Y_j is generated either by the MLR or the MLP;
- GTT_j is converted in an endogenous due date (DD_{en}) as follows: $DD_{en} = t_j + GTT_j$ where t_j is the acceptance time of job j ;
- the endogenous and exogenous due dates are compared:
 - if $DD_{ex} \geq DD_{en}$ the exogenous due date is accepted as is,
 - else, if $DD_{ex} < DD_{en}$ a negotiation starts, and a corrected due date, say CDD , is randomly generated.

Concerning the random generation of CDD, three different probability distributions, representing different negotiation powers, were considered:

- *Balanced market power* When the manufacturer and the customers have, approximately, the same bargain power (or, analogously, the manufacturer has more market power than roughly 50% of its customers, and vice versa), the manufacturer and the customer have the same probability to succeed in the negotiation, getting an advantageous CDD. To reproduce this scenario, CDD is generated, using a uniform distribution on the interval $[DD_{ex}, 1.2DD_{ex}]$, where 1.2 is assumed as the maximum allowance factor that the customer is willing to accept.
- *Manufacturer has more market power* In this case, the manufacturer is more influential than most of its customers, and he will generally win the negotiation. Hence, CDD is generated with a triangular distribution on the interval $[DD_{ex}, 1.2DD_{ex}]$, with modal value located at 3/4 of the interval at $1.15DD_{ex}$.
- *Customer has more market power* In this case, opposite to the previous one, a triangular distribution is used, with modal value located at 1/4 of the interval at $1.05DD_{ex}$.

4.5 Benchmark for the due date generation system

As a benchmark, we also estimated the expected gross throughput time using the approach proposed by Land (2009), which correspond to Eq. (10). We note that the calculation of GTT provided by Eq. (10) is similar to that proposed by Thürer et al. (2019), although the notation is not the same.

$$GTT_{j^*} = \hat{q}_{j^*,p} + M_{j^*} \bar{t} \tag{10}$$

In Eq. (10), for every incoming job j^* , $\hat{q}_{j^*,p}$ is an estimation of the waiting time in the PSP, M_{j^*} is the number of machines visited by j^* , and \bar{t} is the average throughput time (considering both processing and queuing time) at each machine. Due to the symmetry of the system, the same \bar{t} is observed at each machine and, thanks to the stabilization obtained through WLC, this value is almost steady over time.

Owing to these issues, the long-run average of the throughput time, can be effectively used to quantify \bar{t} . In this regard, we note that \bar{t} depends both on the level of

the norms and on the adopted dispatching rule. Yet, in the present study, the value of \bar{t} does not varies much in the four alternative WLC settings that we considered. Hence, the same average value, equal to 4-time units, was used in all four cases.

Conversely, estimating $\hat{q}_{j^*,P}$ is harder because the waiting time in the PSP strongly depends on the current workload situation. Taking the long-run average is not enough and $\hat{q}_{j^*,P}$ must be dynamically estimated at run time through Eq. (11), which is derived from an application of the famous Little’s Law:

$$\hat{q}_{j^*,P} = \max_{m \in R_{j^*}} \left\{ \left(W_m + \sum_{j \in J_P} w_{j,m} \right) - N^* \right\} \tag{11}$$

where R_{j^*} is the routing of an entering job j^* and J_P is the set of jobs currently pending in the PSP.

Note that, when job j^* enters the PSP, it will wait there on average until all the norms drop below the common threshold value N^* . For this reason Eq. (11) quantifies the waiting time in the PSP as the maximum time-gap between the threshold N^* and the workload W_m released to machine m , added to the workload $\left(\sum_{j \in J_P} w_{j,m} \right)$ that will be released to m in the next future.

As noted above, $\hat{q}_{j^*,P}$ varies over time, but just to give an idea, its average value was equal to 7 units of time.

4.6 Obtained results

The results we obtained are shown in Tables 5 and 6. The tables report the results for each investigated scenario, and the scenario details are indicated in the first three columns: (i) the dispatching rule, which is EDD in Table 5 and ODD in Table 6, (ii) the value of the optimized norms, as explained in Fig. 2 and Table 3, and the bargain power (cf. Sect. 4.4). For each of these scenarios, the results report the percentage of tardy jobs (columns 4–8), and the percentage of times that a negotiation was started (columns 9–11). Displayed values are averaged over ten simulation runs, made for each combination of dispatching rule, optimized norm level, negotiation power, and system type. As additional benchmarks, both tables also display the percentage of tardy jobs observed for the Standard Push and for the Standard WLC systems, where the term standard means operating without a forecasting and negotiation system (columns 4–5).

As it can be seen, WLC coupled with an effective forecasting system consistently outperform all the other configurations; the performances of the models are statistically significant, as it is confirmed by ANOVA (Tables 7 and 8) and pairwise Bonferroni post hoc test (Table 9), with a P value of 0.01.

First, we note that, as expected, the percentage of tardy jobs and the number of performed negotiations are persistently lower when the ODD dispatching rule is used to sort jobs in the PSP and in the shop floor. This result is in line with much WLC literature that confirmed the capability of the ODD rule to stabilize the system, reducing WIP and the GTT variability (Thürer et al. 2019). However, a remarkable

Table 5 Percentage of tardy jobs and negotiated DDs with EDD dispatching rule

Dispatch. Rule	Norm level	Bargain power	% of tardy jobs			% of negotiated DD					
			GTT is not forecasted			GTT is forecasted			GTT is forecasted		
			Standard push	Standard WLC	Standard WLC	WLC+LAND	WLC+MLR	WLC+MLP	WLC+LAND	WLC+MLR	WLC+MLP
EDD PSP & shop floor	$N^*=5$	Customer	14%	13.80%	13.19%	12.17%	12.43%	19.65%	28.13%	30.27%	
	Opt. WIP	Manufacturer	14%	13.80%	12.86%	11.80%	11.58%	19.03%	27.45%	27.60%	
		Balanced	14%	13.80%	12.93%	11.91%	11.77%	19.98%	31.00%	27.51%	
	$N^*=10$	Customer	14%	7.30%	8.00%	5.62%	5.39%	18.82%	26.62%	19.69%	
	Opt. Tardy	Manufacturer	14%	7.30%	7.41%	4.20%	5.08%	18.52%	21.81%	20.10%	
		Balanced	14%	7.30%	7.70%	4.95%	4.48%	19.12%	24.85%	17.23%	

Table 6 Percentage of tardy jobs and negotiated DDs with ODD dispatching rule

Dispatch. rule	Norm level	Bargain power	% of tardy jobs			% of negotiated DD					
			GTT is not forecasted			GTT is forecasted			GTT is forecasted		
			Standard Push	Standard WLC	Standard WLC	WLC+LAND	WLC+MLR	WLC+MLP	WLC+LAND	WLC+MLR	WLC+MLP
ODD PSP & shop floor	N* = 5	Customer	11.0%	7.70%	7.70%	7.58%	7.22%	7.40%	17.53%	1.56%	0.55%
	Opt. WIP	Manufacturer	11.0%	7.70%	7.70%	7.57%	7.52%	7.31%	17.65%	1.72%	0.56%
		Balanced	11.0%	7.70%	7.70%	7.58%	7.29%	7.44%	17.20%	1.51%	0.55%
N* = 6	Opt. Tardy	Customer	11.0%	6.56%	6.56%	6.77%	6.06%	5.92%	16.62%	2.51%	0.74%
		Manufacturer	11.0%	6.56%	6.56%	6.68%	5.80%	6.01%	16.34%	1.88%	0.66%
		Balanced	11.0%	6.56%	6.56%	6.78%	5.80%	5.96%	16.12%	1.63%	0.60%

Table 7 ANOVA report—assessing mean values and variance of percentage of tardy jobs in 1000 observations with different PPC and forecasting systems

Groups	Number of observations	Sums of values	Mean % tardy	Variance of % tardy
Standard push—no forecasting system	1000	146.018	0.146	0.008
Standard WLC—no forecasting system	1000	69.230	0.069	0.002
WLC—land forecasting system	1000	66.149	0.066	0.002
WLC—MLR forecasting system	1000	53.671	0.054	0.001
WLC—MLP forecasting system	1000	56.592	0.057	0.001

Table 8 ANOVA results—assessing the statistical significance of variance between different PPC and forecasting systems

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F value	P value	F-Crit.
Between groups	5.893	4	1.473	563.790	0	2.4
Within groups	13.053	4995	0.003			
Total	18.947	4999				

Table 9 P values of the pairwise Bonferroni post hoc test

	Standard WLC	WLC + Land	WLC + LR	WLC + MLP
Standard push	2.387 E-111	4.376 E-119	3.992 E-155	5.400 E-147
Standard WLC		8.690 E-2	9.100 E-23	6.343 E-15
WLC + Land			1.347 E-15	2.363 E-09
WLC + LR				2.713 E-02

and unexpected exception must be noted. Indeed, regardless of the market power of the manufacturer, for a WLC system with norms optimized to minimize the percentage of tardy jobs, the best results are obtained using the EDD, rather than the ODD dispatching rule. In particular, the overall best result of 4.20% tardy jobs is observed for the configuration with: EDD dispatching rule, MLR forecasting model, higher negotiation power of the manufacturer. Conversely, the ODD cannot go below a 5.80% of tardy jobs, a result obtained with the MLR forecasting system and balanced or manufacturer-side negotiation power. This interesting result demonstrates that even a simple dispatching rule, very common and easy to be implemented in the industry, can be effectively used if WLC is supported by an accurate forecasting system. Also, and perhaps more important, this result proves the accuracy of the forecasting systems that, according to the current system state, succeeds in identifying the critical jobs, for which an extension of the due date is necessary.

Honestly, we must note that the interesting performances in terms of on-time delivery of the EDD system are obtained by bargaining the DD around 20% of times

(i.e. values in columns 9–11 of Table 4), whereas the results of the ODD system are obtained with a very low number of negotiations, which are activated about 1% of times, and consistently below 1% if the MLP is used to forecast the GTT (i.e. values in columns 11 of Table 5). Such reduction in the number of negotiations is due to the decrease and to the higher stabilization of the gross throughput time obtained when the EDD is replaced by the ODD. Indeed, as Table 3 shows, GTT is reduced by more than 10% (from 25.8 to 23.1-time units), thus reducing the probability that the exogenous DDs could be considered too tight. Anyhow, a level of negotiation of about 20% does not seem excessive in industrial cases, also because customers are generally inclined to accept (slightly) higher DDs, if they are more reliable.

A last remark must be made regarding the percentage of negotiated due dates displayed in Tables 5 and 6. When forecasts are made via MLR or MLP, this percentage changes depending both on the simulated scenario and on the market power. This fact can be explained as follows. Both MLR and MLP forecast GTT based on the current system's state, but any time a due date is extended, also the system's state gets modified, because the jobs pending in the PSP and, subsequently, those ones queuing in the shop floor will be sorted in a different way. In other words, forecasts depend on the system state which, in turn, is influenced by the forecasts; it is exactly this 'cyclic link' that explains the observed change in the percentage of contracted due dates.

Conversely, this effect is less marked when forecasts are made according to the approach proposed by Land. Indeed, Eq. (10) has both a constant ($M_{j^*} \bar{t}$) and a dynamic part ($\hat{q}_{j^*,P}$). Since the latter one is smaller and it only depends on the current state of the PSP (and not of the whole system), the percentage of contracted DDs does not change much and oscillates around 16–19% in each observed scenario.

5 Further managerial implications

In this section, we report some insight on the GTT forecasting and negotiation method. In particular, Sect. 5.1 investigates whether two simple negotiation methods applied to standard push and standard WLC, namely the blind and the selective negotiation methods, can provide similar results compared to those of Sect. 4.6. In Sect. 5.2, we introduce the reverse negotiation, thus giving the manufacturer the possibility to propose matter-of-factly closer DDs to the customer, with the aim of enhancing customer satisfaction and loyalty.

5.1 Deepening the results: blind and selective negotiation with standard push and standard WLC systems

To further investigate the obtained results, we performed an additional set of simulations, aiming to test whether the increased number of on time deliveries is really due to the deployment of the forecasting system, or if it is just the effect of the negotiation, and of the consequent extension of the due dates. Indeed, stretching the reasoning to an extreme, a forecasting system that persistently

predicts very long GTTs would trigger the negotiation process almost 100% of the times. In this case, therefore, due dates would be generated on the enlarged interval [28, 67.2], where the upper limit (67.2) corresponds to the original one (52) multiplied for the maximum allowable extension of 1.2.

To test this hypothesis, we added a ‘blind’ negotiation procedure, both to the Standard Push and Standard WLC systems, i.e. non supported by the forecasting system. More precisely, we introduced the possibility to negotiate the due dates in two different ways:

- *Blind negotiation* In this case, 20% of the jobs are randomly selected, and their due date is negotiated. The value of 20% was chosen to assure a fair comparison, as this percentage closely matches the one observed when the forecasting system were used (see Table 5 for details).
- *Selective negotiation* In this case, the negotiation is limited to the jobs with a tight due date. Specifically, 50% of the jobs with a due date shorter than 39.2 time-units are randomly selected and their due date is negotiated. Since 39.2 corresponds to 2/5 of the interval [28, 56], on which due dates are generated, also in this case we have a total of 20% negotiated due dates.

Since the EDD was proven to be very effective in reducing the percentage of tardy job (provided that a percentage of negotiations around 20% is admissible), the analysis was limited to this dispatching rule, relatively to the balanced market power case. Specifically, for the Standard Push system, we found an average percentage of tardy jobs of 13.8% and 14.2% in case of blind and selective negotiation, respectively. Surprisingly, both values are perfectly aligned with the percentage of tardy jobs (14%) of the original push system, as confirmed by a t-test (for means comparison), which resulted negative at level $\alpha=0.05$. Although this lack of improvement might seem counterintuitive, it can be explained as follows. Since jobs are sorted based on their DD, an extension of the DD results in a reduction of priority both within the PSP and in the queues before every machine. Such priority reduction increases the total waiting time, thereby cancelling the benefit of the extra allowance of DD.

The situation is partially subverted using WLC. In this case, due to variability reduction, the negotiation procedure has a positive effect and the percentage of tardy jobs drops from the original value of 7.3% to 6.3% and to 6.5% in case of blind and selective negotiation, respectively. These reductions are statistically significant (at level $\alpha=0.05$), but the observed values remain significantly higher than the minimum percentage of tardy jobs that were achieved using either the MLR or the MLP forecasting system (cf. Table 5). We believe that this result is an additional proof of the quality and robustness of the DDs’ generation system we proposed. Indeed, it confirms the importance of WLC to stabilize the system, and of an accurate forecasting system that, based on the current system state, makes it possible to identify the critical jobs that really need a due date extension.

5.2 Reverse negotiation: offering earlier DDs to enhance customer's loyalty

In the previous sections, the negotiation procedure was implicitly considered as a defensive approach, used by the manufacturer to postpone the exogenous DDs considered too hard to be met. In other words, a negotiation started any time the exogenous due date DD_{ex} was earlier than the endogenous one DD_{en} . Conversely, in the opposite case, the exogenous due date was passively accepted, even if it was very late, i.e. if $DD_{en} \ll DD_{ex}$. In this case, however, it could be wise for the manufacturer to use, to his advantage, the time gap $\Delta D = (DD_{ex} - DD_{en})$. For instance, the manufacturer could offer to his customer an earlier DD and this reduction could probably enhance the satisfaction and the loyalty of the customer, or it might even allow an extra profit.

To assess the viability of this 'reverse' negotiation scheme, we modified the negotiation process described in Sect. 4.4 in the following way:

- If $DD_{ex} < DD_{en}$ the standard negotiation starts,
- Else, $DD_{ex} \geq DD_{en}$:
 - If $\Delta D > (\alpha \cdot DD_{en})$ then a corrected due date $CDD = (1 + \alpha) \cdot DD_{en}$ is proposed to the customer,
 - If $\Delta D \leq (\alpha \cdot DD_{en})$ then the exogenous due date is accepted as is i.e. $CDD \equiv DD_{ex}$

where α is a safety coefficient, dependent on the risk aversion of the manufacturer. Clearly, the lower the value of α , the more aggressive and riskier is the reverse negotiation policy.

Repeating the same simulation runs described in the previous section, limited to the case of balanced bargain power and EDD dispatching rule, the results of Table 10 were finally obtained.

Inevitably, the introduction of reverse negotiation leads to a deterioration in the observed performance and, obviously, the worsening is more pronounced for low value of the safety coefficient. Nonetheless, the deterioration is very limited (less than one percentage point) and, most of all, even at a level $\alpha=0.3$, the percentage of tardy jobs remains lower than that observed for the standard WLC system without forecasting system. This makes the reverse negotiation scheme very attractive both from an operating and managerial point of view.

5.3 Stress test: measuring the predictive power in highly saturated systems

To conclude the analysis, we report a further investigation that we made to assess the impact, on the observed performance, of the utilization level of the system. To this aim, we altered the simulation environment to increase the utilization level (of each machine) from 90 to 95%, an operating condition that frequently occurs in make to order job shops. Specifically:

Table 10 Percentage of tardy jobs with EDD rule, Balanced bargain power and different negotiation procedures

Norms level	Alfa level	% of tardy jobs														
		GTT is not forecasted					Forecasted GTT & standard negotiation					Forecasted GTT & reverse negotiation				
		Standard Push	Standard WLC	WLC & LAND	WLC & MLR	WLC & MLP	WLC & LAND	WLC & MLR	WLC & MLP	WLC & LAND	WLC & MLR	WLC & MLP	WLC & LAND	WLC & MLR	WLC & MLP	
N ^r = 5 Opt. WIP	$\alpha=0.3$	14.00%	13.80%	12.93%	11.91%	11.77%	14.80%	12.06%	13.85%							
	$\alpha=0.5$	14.00%	13.80%	12.93%	11.91%	11.77%	13.88%	11.20%	11.80%							
	$\alpha=0.6$	14.00%	13.80%	12.93%	11.91%	11.77%	14.43%	11.48%	12.80%							
N ^r = 10 Opt. Tardy	$\alpha=0.3$	14.00%	7.30%	7.70%	4.95%	4.48%	7.48%	5.90%	6.00%							
	$\alpha=0.5$	14.00%	7.30%	7.70%	4.95%	4.48%	6.71%	5.20%	5.55%							
	$\alpha=0.7$	14.00%	7.30%	7.70%	4.95%	4.48%	6.68%	5.25%	4.53%							

- to achieve the desired utilization level, the arrival rate λ has been increased to 1.64 jobs per unit of time;
- the investigation has been limited to the scenario with the EDD dispatching rule;
- since the Poisson process modelling jobs' arrival has been changed, the optimization procedure (needed to find the optimal level of the norms) has been repeated and an optimal norm level equal to 6 was found.

Relatively to the last point, it is worth noting that, due to the high utilisation level herein considered, the optimal level of the norms that we found minimize not only the tardiness, but also the WIP level, thus assuring a double economic benefit. The other performance (expressed in terms of % of tardy jobs and % of negotiated DDs) are shown in Table 11.

As it can be seen, when a forecasting system is not used, the percentage of tardy jobs is two to four time higher (for the Push and WLC system, respectively) than that of the WLC system with forecasted GTT. Obviously, due to the increase of the average utilization (the productive capacity of the system is almost saturated and, due to the variability of the Poisson generating process, occasionally demand may be even higher than capacity) the % of tardy jobs has increased. Yet, if we compare the results of Table 5, relatively to the WIP minimization case, the worsening is not so bad, approximatively 1.5 times.

These results confirm the quality and robustness of our approach, which can be fruitfully integrated also in more realistic scenarios, hence strengthening the operational appeal of our work.

6 Conclusions

The paper focused on WLC and showed that this production planning and control technique is ideal to maximize on time deliveries, especially if it is coupled with a robust and consistent forecasting system, aimed to estimate the GTT of the accepted jobs and to define reliable delivery dates. To build the forecasting system, we propose regressing the GTT using as explanatory variables the workloads of the jobs pending in the PSP and that of the jobs already released to the shop floor. Due to the non-linearity among dependent and independent variables, forecasts were made using a multi-layer perceptron (MLP); however, to have a benchmark solution, a linear and a quadratic regression model were also developed.

The model was tested in a pure job-shop, with six equally loaded machines (at 90 and 95% utilisation levels), reproduced in a simulative environment. Also, to increase the realism of the simulation, rather than generating due dates in a purely random way, as it is rather common in WLC literature, we introduced a negotiation scheme, supported by our GTT forecasting model. More precisely, if the external due date is too tight, relatively to the estimated GTT, a negotiation starts, and the manufacturer tries to extend the exogenous due date. Obtained results are very promising, as they demonstrate the quality and the robustness of our approach in identifying the critical jobs that really need an adjustment of the due date, i.e. its postponement. In this respect, the MLP offers superior performance, as it systematically cuts down the

Table 11 Percentage of tardy jobs with 95% of saturation level

% of tardy jobs		% of negotiated DDs			
GTT is not forecasted		GTT is forecasted			
Standard push	Standard WLC	WLC+LAND	WLC+MLR	WLC+MLP	WLC+MLR
60.44%	23.24%	20.66%	19.72%	19.77%	72.86%
					62.82%
					62.79%

percentage of tardy jobs with a low percentage of negotiated DDs. Nonetheless, also the linear multi regression model performs very well, as its performance, expressed as the percentage of tardy jobs, always outperform that of an equivalent WLC system unsupported by a forecasting method. To further demonstrate the operating and managerial implications of our methods, we also showed that the forecasting system makes it possible to adopt an interesting managerial policy that we named 'reverse negotiation scheme'. Specifically, when the exogenous due dates are long, the manufacturer can offer a shortened due date to the customer, aiming to increase its satisfaction and loyalty, or to get an extra profit. This policy was shown to be very attractive as the very limited increase of tardy jobs might be significantly offset by the number of orders delivered in advance. In general, therefore, our method has several managerial implications, as it assures a real competitive edge in terms of increased customer satisfaction and loyalty and smoother production. In this regard, we finally note that an industrial implementation could be straightforward and relatively cheap, as the forecasting model can be developed easily leveraging open source libraries and analysing data gathered from a Manufacturing Execution System. Whether data already stored would prove to be not sufficient, a digital twin could be leveraged to reproduce the manufacturing system by simulation, with the aim of gathering additional data for the training phase.

The paper also showed that, as expected, to maximise the number of on time deliveries, before deploying the forecasting system WLC norms should be fine-tuned to minimize the percentage of tardy jobs. Nonetheless, the forecasting system makes WIP minimisation an alternative option. In WLC literature, WIP minimisation is rarely considered, as it makes the jobs' release phase very restrictive, leading to high waiting time in the PSP and to a higher probability of a job to go late. However, thanks to the forecasting system, this negative effect can be counterbalanced and, indeed, it is possible to obtain a percentage of tardy jobs lower than that of an analogous purely push operating system. Certainly, to further investigate this possibility, in addition to the percentage of tardy job, additional criteria should be considered, such as stockholding costs and penalty costs due to early and/or late deliveries. This could be an interesting topic for future works.

Other improvements could regard a refinement of the forecasting models. As a first attempt, alternative ways to improve the performance of the regression should be investigated. In this regard, an interesting possibility could be that to use a regression model based on linear basis functions that have the property to be linear function of the parameter and yet can be nonlinear with respect to the input variables. Also, another interesting opportunity could be that of trying some regularization methods for the regression models, such as lasso (least absolute shrinkage and selection operator). Indeed, a method that executes both variable selection and regularization could enhance the prediction accuracy of the model, without the need of expert intuition or time-consuming activities of feature selection.

Moreover, to reduce the forecast error of both the regression and the neural network, it could be useful to add additional explanatory variables, in order to estimate the waiting time in the PSP and the waiting time on the shop floor, separately. It could be even possible to estimate the waiting time for each machine queue, as this would assure a more detailed control of the system, offering additional levers of

action to the production manager. Lastly, the estimate of the gross and/or of the shop floor throughput time could be used not only to trigger the bargaining system, but also to define new and more effective dispatching rules, so as to give greater priority to the jobs for which the due date extension is deemed insufficient.

Funding Open access funding provided by Università degli Studi di Parma within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akillioglu H, Dias-Ferreira J, Onori M (2016) Characterization of continuous precise workload control and analysis of idleness penalty. *Comput Ind Eng* 102(December 2016):351–358. <https://doi.org/10.1016/j.cie.2016.06.026>
- Bechte W (1988) Theory and practice of load-oriented manufacturing control. *Int J Prod Res* 26(3):375–395
- Bechte W (1994) Load-oriented manufacturing control just-in-time production for job shops. *Prod. Plan. Control* 5(3):292–307. <https://doi.org/10.1080/09537289408919499>
- Bergamaschi D, Cigolini R, Perona M, Portioli A (1997) Order review and release strategies in a job shop environment: a review and a classification. *Int J Prod Res* 35(2):399–420. <https://doi.org/10.1080/002075497195821>
- Bertolini, M., Romagnoli, G., Zammori, F.: Assessing performance of Workload Control in High Variety Low Volumes MTO job shops: A simulative analysis. In: Proceedings of 2015 International Conference on Industrial Engineering and Systems Management, IEEE IESM 2015, pp. 362–370 (2016a). <https://doi.org/10.1109/IESM.2015.7380184>
- Bertolini, M., Romagnoli, G., Zammori, F.: Simulation of two hybrid production planning and control systems: a comparative analysis. In: Proceedings of 2015 International Conference on Industrial Engineering and Systems Management, IEEE IESM 2015, pp. 388–397 (2016b). <https://doi.org/10.1109/IESM.2015.7380187>
- Bertolini M, Braglia M, Romagnoli G, Zammori F (2013) Extending value stream mapping: the synchronous MRP case. *Int J Prod Res* 51(18):5499–5519. <https://doi.org/10.1080/00207543.2013.784415>
- Bertolini M, Romagnoli G, Zammori F (2017) 2MTO, a new mapping tool to achieve lean benefits in high-variety low-volume job shops. *Prod. Plan. Control* 28(5):444–458. <https://doi.org/10.1080/09537287.2017.1302615>
- Bertrand JWM (1981) The effect of Workload Control on order flow times. In: Brans JP (ed) *Operational Research '81*. North Holland Publishing Company, Amsterdam, The Netherlands, pp 779–790
- Bertrand J, Van Ooijen H (2002) Workload based order release and productivity: a missing link. *Prod. Plan. Control* 13(7):665–678. <https://doi.org/10.1080/0953728021000026276>
- Bhosale KC, Pawar PJ (2019) Material flow optimisation of production planning and scheduling problem in flexible manufacturing system by real coded genetic algorithm (RCGA). *Flex. Serv. Manuf. J.* 31(2):381–423. <https://doi.org/10.1007/s10696-018-9310-5>
- Chiang TC, Fu LC (2007) Using dispatching rules for job shop scheduling with due date-based objectives. *Int J Prod Res* 45(14):3245–3262. <https://doi.org/10.1080/00207540600786715>

- Cigolini R, Portioli-Staudacher A (2002) An experimental investigation on workload limiting methods with ORR policies in a job shop environment. *Prod. Plan. Control* 13(7):602–613
- Dörner J, Günther HO, Gujjala R (2013) Master production scheduling and sequencing at mixed-model assembly lines in the automotive industry. *Flexible Serv. Manuf. J.* 27(1):1–29. <https://doi.org/10.1007/s10696-013-9173-8>
- Fernandes NO, Land MJ, Carmo-Silva S (2014) Workload control in unbalanced job shops. *Int J Prod Res* 52(3):679–690. <https://doi.org/10.1080/00207543.2013.827808>
- Fernandes NO, Land MJ, Carmo-Silva S (2016) Aligning workload control theory and practice: lot splitting and operation overlapping issues. *Int J Prod Res* 54(10):2965–2975. <https://doi.org/10.1080/00207543.2016.1143134>
- Fernandes NO, Thürer M, Silva C, Carmo-Silva S (2017) Improving workload control order release: incorporating a starvation avoidance trigger into continuous release. *Int J Prod Econ* 194(December 2017):181–189. <https://doi.org/10.1016/j.ijpe.2016.12.029>
- Germis R, Riezebos J (2010) Workload balancing capability of pull systems in MTO production. *Int J Prod Res* 48(8):2345–2360. <https://doi.org/10.1080/00207540902814314>
- Géron A (2019) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* (Oreilly & Associates Inc, Ed.) O'Reilly Media, Sebastopol
- Harrod S, Kanet J (2013) Applying work flow control in make-to-order job shops. *Int J Prod Econ* 143(2):620–626. <https://doi.org/10.1016/j.ijpe.2012.02.017>
- Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical. Learning.* <https://doi.org/10.1007/978-0-387-84858-7>
- Hendry L, Huang Y, Stevenson M (2013) Workload control: successful implementation taking a contingency-based view of production planning & control. *Int. J. Oper. Prod. Manag.* 33(1):69–103. <https://doi.org/10.1108/01443571311288057>
- Henrich P, Land MJ, Gaalman GJC (2007) Semi-interchangeable machines: implications for workload control. *Prod. Plan. Control* 18(2):91–104. <https://doi.org/10.1080/09537280600909445>
- Hopp WJ, Spearman ML (2008) *Factory Physics*, 3rd edn. McGraw-Hill/Irwin, Dubuque, Iowa
- Kingma, D. P., Ba, J.: *Adam: A Method for Stochastic Optimization.* <http://arxiv.org/abs/1412.6980> (2014)
- Land MJ (2006) Parameters and sensitivity in workload control. *Int J Prod Econ* 104(2):625–638. <https://doi.org/10.1016/j.ijpe.2005.03.001>
- Land MJ (2009) Cobacabana (control of balance by card-based navigation): a card-based system for job shop control. *Int J Prod Econ* 117(1):97–103. <https://doi.org/10.1016/j.ijpe.2008.08.057>
- Land MJ, Gaalman GJC (1996) Workload control concepts in job shops. A critical assessment. *Int J Prod Econ* 46–47:535–548
- Land MJ, Gaalman GJC (1998) The performance of workload control concepts in job shops: improving the release method. *Int J Prod Econ* 56–7:347–364
- Little JDC (1961) A proof for the queuing formula. *Oper Res* 9(3):383–387
- Marangoni, G., Romagnoli, G., Zammori, F.: Multiple value stream mapping: how to implement workload control in complex systems. In: *IESM'2013, international conference on industrial engineering and systems management*, Rabat, Morocco (2013)
- Mezzogori D, Romagnoli G, Zammori F (2019) Deep learning and WLC: how to set realistic delivery dates in high variety manufacturing systems. *IFAC-PapersOnLine* 52(13):2092–2097. <https://doi.org/10.1016/j.ifacol.2019.11.514>
- Miller BI (1969) A queueing reward system with several customer classes. *Manag. Sci.* 16:234–245
- Montavon G, Samek W, Müller K-R (2018) Methods for interpreting and understanding deep neural networks. *Digit Signal Proc* 73:1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
- Moreira MRA, Alves RAFS (2009) A methodology for planning and controlling workload in a job-shop: a four-way decision-making problem. *Int J Prod Res* 47(10):2805–2821. <https://doi.org/10.1080/00207540701725083>
- Moreira, M. R., Alves, R.: A new input-output control order release mechanism: how workload control improves manufacturing operations in a job shop. In *Faculdade De Economia Working Papers, Research Paper Series*, vol. 107 (2006a)
- Moreira, M. R., Alves, R.: Does order negotiation improve the job-shop workload control? In: *EurOMA 2006 Conference Proceedings*, Glasgow, Scotland, No. 18, pp. 741–749 (2006b)
- Nandi, A., & Rogers, P.: Behavior of an order release mechanism in a make-to-order manufacturing system with selected order acceptance. In: *Chick, S., Sánchez, P., Ferrin, D., Morrice, D. (eds.)*

- Proceedings of the 2003 Winter Simulation Conference, vol. 2, pp. 1251–1259. <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1261558> (2003)
- Nandi, A.: Input control strategies for make-to-order manufacturing systems via order acceptance/rejection. University of Calgary (Alberta) (2000)
- Oosterman B, Land MJ, Gaalman GJC (2000) The influence of shop characteristics on workload control. *Int J Prod Econ* 68(1):107–119
- Scott M (1970) Queueing with control on the arrival of certain types of customers. *CORS J.* 8:75–86
- Stevenson M (2006) Refining a workload control (WLC) concept: a case study. *Int J Prod Res* 44(4):767–790. <https://doi.org/10.1080/00207540500338070>
- Stevenson M, Hendry LC, Kingsman BG (2005) A review of production planning and control: the applicability of key concepts to the make-to-order industry. *Int J Prod Res* 43(5):869–898. <https://doi.org/10.1080/0020754042000298520>
- Stevenson M, Huang Y, Hendry LC, Soepenber E (2011) The theory and practice of workload control: a research agenda and implementation strategy. *Int J Prod Econ* 131(2):689–700. <https://doi.org/10.1016/j.ijpe.2011.02.018>
- Thürer M, Stevenson M (2016a) Card-based delivery date promising in pure flow shops with order release control. *Int J Prod Res* 54(22):6798–6811. <https://doi.org/10.1080/00207543.2016.1177672>
- Thürer M, Stevenson M (2016b) Workload control in job shops with re-entrant flows: an assessment by simulation. *Int J Prod Res* 54(17):5136–5150. <https://doi.org/10.1080/00207543.2016.1156182>
- Thurer M, Silva C, Stevenson M (2011) Optimising workload norms: the influence of shop floor characteristics on setting workload norms for the workload control concept. *Int J Prod Res* 49(4):1151–1171. <https://doi.org/10.1080/00207541003604836>
- Thürer M, Stevenson M, Silva C (2011) Three decades of workload control research: a systematic review of the literature. *Int J Prod Res* 49(23):6905–6935. <https://doi.org/10.1080/00207543.2010.519000>
- Thürer M, Stevenson M, Silva C, Land MJ, Fredendall LD, Melnyk SA (2014) Lean control for make-to-order companies: integrating customer enquiry management and order release. *Prod. Oper. Manag.* 23(3):463–476. <https://doi.org/10.1111/poms.12058>
- Thürer M, Land MJ, Stevenson M, Fredendall LD (2017) On the integration of due date setting and order release control. *Prod. Plan. Control* 28(5):420–430. <https://doi.org/10.1080/09537287.2017.1302102>
- Thürer M, Stevenson M, Land MJ, Fredendall LD (2019) On the combined effect of due date setting, order release, and output control: an assessment by simulation. *Int J Prod Res* 57(6):1741–1755. <https://doi.org/10.1080/00207543.2018.1504250>
- Tikk D, Kóczy LT, Gedeon TD (2003) A survey on universal approximation and its limits in soft computing techniques. *Int J Approx Reason* 33(2):185–202
- van Ooijen HPG, Bertrand JWM (2001) Economic due-date setting in job-shops based on routing and workload dependent flow time distribution functions. *Int J Prod Econ* 74(1–3):261–268. [https://doi.org/10.1016/S0925-5273\(01\)00131-1](https://doi.org/10.1016/S0925-5273(01)00131-1)
- Wiendhal H-P (1995) *Load-Oriented Manufacturing Control*. Springer, Berlin
- Yadav V, Jain R, Mittal ML, Panwar A, Lyons AC (2019) The propagation of lean thinking in SMEs. *Prod. Plan. Control* 30(10–12):854–865. <https://doi.org/10.1080/09537287.2019.1582094>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Davide Mezzogori graduated with distinction in 2015 in Management Engineering and received a Ph.D. in industrial engineering in 2019 at the University of Parma (Italy). From 2019 he is a post-graduate research fellow at the Department of Architecture and Engineering of the University of Parma, where he is involved in the DigiLab4U international project (<http://digilab4u.com/>), for the development of a serious game for supply chain and operation management education. His main research interests concern the application of Machine and Deep Learning algorithms to Industrial problems, as well as the development of metaheuristics for operation management. He applied these techniques both academically and industrially; obtained results were published on prestigious International Journals and/or presented at International Conferences.

Giovanni Romagnoli was born in November 1982 in Scandiano (RE), Italy, the town where he still lives with his family. He is a father-of-three. He received his Master Degree (with honours) in Mechanical Engineering for the Food Industry at the University of Parma, Italy. In 2013, he completed his PhD course in Operations and Project Management at the Department of Engineering and Architecture of Parma, defending the thesis entitled "Hybrid production planning and control systems: towards an application of lean manufacturing to the Make-To-Order sector". Since 2013, Giovanni Romagnoli is employed as Research Fellow at the same Department. Since December 2018, he acts as a local manager on the project DigiLab4U (<http://digilab4u.com/>). His research interests include RFID and supply chain management, production planning & control systems, improvements and applications of lean manufacturing, designing and managing food plants and processes; and led to the realization of more than 40 works published on International Journals or presented at International Conferences.

Francesco Zammori graduated with distinction in 2004 in Management Engineering and completed his postgraduate studies in 2009, when he received a PhD in Industrial Engineering from the University of Pisa. From 2012, he works at the University of Parma (Italy), currently as associate professor, where he teaches Management accounting systems, Data science for business management, Databases and Information systems. His research interests mainly concern: (i) Lean Manufacturing, (ii) Hybrid Production Planning and Control Systems, (iii) Modelling and Simulation and (iv) Statistics and Machine Learning. His research activities led to the realisation of more than 50 scientific works, 40 of which have been published by prestigious International Journals.