2022 IEEE 18th International Conference on Automation
Science and Engineering (CASE)
August 20-24, 2022. Mexico City, Mexico

2022 IEEE 18th International Conference on Automation Science and Engineering (CASE) | 978-1-6654-9042-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CASE49997.2022.9926570

A Novel Approach to Modeling of Production System: A Case Study at a Small/medium-sized Manufacturer

Yuting Sun and Liang Zhang

Abstract— **High-fidelity mathematical models are essential to implement model-based analysis and control in manufacturing research and practice. Currently, such models are typically conducted manually in an ad hoc manner. This approach presents several limitations, especially to small and medium-sized manufacturers, such as unavailability of equipment status data, inconvenient data collection process, non-standard and non-unique modeling rules, etc. In this paper, we describe a case study at a local small manufacturer of medical devices and apply a novel approach of production system modeling to overcome various practical challenges in collecting up- and downtime data of the operations. Specifically, the parametric model of the production system is identified based on system performance metrics derived from the parts flow data. With the model constructed, system bottleneck is analyzed and then, to enhance system throughput, potential improvement actions including operation speed-up, downtime reduction, and buffer expansion are explored. Finally, model sensitivity is analyzed by comparing the deviation of the model-predicted performance metrics to those produced by a reference nominal model.**

I. Introduction

The emergence and development of Industry 4.0 during the past decade have lead to great advances in computation power and data analytics technologies. Data-driven modeling and analysis has since become increasingly popular in manufacturing research and practice. However, small/medium-sized manufacturers (SMMs) typically face various challenges in utilizing the production data to create mathematical models for their production systems to improve their production processes. In current manufacturing research, a production system is typically modeled as a stochastic process, where the operations/machines are characterized by randomly distributed uptimes, downtimes, and cycle times [1]–[3]. The conventional approach to determine the parametric system model is to collect the status data (i.e., up- and downtimes) from each individual workstation and extract the model parameters by aggregating and analyzing those data. To apply this approach, one of the greatest challenges for the SMMs is the unavailability of system-wide automatic data collection technologies or equipment that can monitor the status of all operations/machines and record the data in real-time. Moreover, even when machine status data are available (either through automatic or manual collection), the data cleaning process is often tedious and complicated due to issues such as noise, data integrity and ambiguity.

*This work was supported in part by the U.S. National Science Foundation (NSF), under Grant Number FM-2134367.

Yuting Sun and Liang Zhang are with Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, 06269, USA. Email: yuting.2.sun@uconn.edu, liang.zhang@uconn.edu

As reported in the literature, the conventional production system modeling approach often leads to significant efforts in collecting, cleaning, and processing the production data and in designing algorithms to calculate the system parameters. For instance, in the case study of an automated pizza production line carried out in [4], the average downtime of each machine is computed based on hand-written records of failures spanning a period of four years in order to present a statistical analysis of the failure data and investigate the improvement of this production line. Moreover, in [5], the authors collect the up-/downtime data from 15 machines to calculate the parameters of a two-stage aggregated line model simplified from a complex motorcycle manufacturing system, which involves a large amount of data processing work. In [6], a Bernoulli model is built for a production system with quality control devices in a picture tube plant. A large amount of data of up-/downtime of each machine are collected to compute the Bernoulli parameters. The similar method is also adopted by [7] in a case study at an automotive body shop. Besides, in paper [8], a method to model a manufacturing system for productivity improvement is introduced and a software tool is developed to implement this method by connecting the analytical algorithm with the automatic data acquisition system for real-time production data collection. This method, however, still requires a customized calculation algorithm for the system considered. Clearly, the lack the in-house expertise, extra workforce, and IT infrastructure makes it it difficult for the SMMs to effectively apply this production modeling approach in their practice.

To overcome these limitations, it is proposed in [9]–[11] to reversely compute the parameters of a production system model based on measured system performance metrics, i.e., through inverse modeling [12], [13]. In this new approach, the input are the system performance metrics (e.g., throughput, work-in-process), which have commonly accepted definitions. This can greatly reduce the ambiguity that may be contained in operating status data. Moreover, the performance metrics used in this approach can be measured based on part-counting. This can be easily accomplished by deploying sensors (e.g., weight sensors, photoelectric sensors) into the manufacturing process. Thus, the goal of this paper is to extend this new modeling approach to an asynchronous exponential production system models and apply and test the robustness of this approach in a case study.

The rest of the paper is organized as follows: The system operation and problems addressed are described in Section II. Mathematical modeling is implemented in Section III. Section IV analyzes the baseline system performance and

explores some improvement actions for increasing the system throughput. The model sensitivity is discussed in Section V. Finally, the conclusions and future work are summarized in Section VI.

II. System Operation and Problem Description

A. System layout and operation

The production system studied in this paper is from a local SMM, which designs and produces electrical/mechanical devices for medical applications. The layout of this particular production system is shown in Fig. 1.

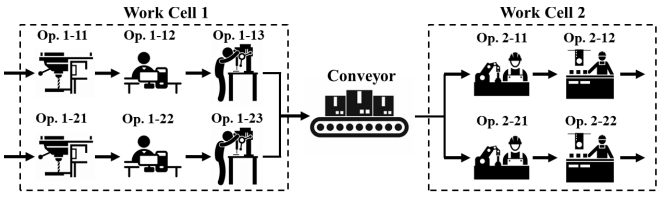

Fig. 1. Layout of the medical device production system

There are two work cells in this production system. Both work cells consist of two parallel lines, which split the work load and carry out identical processing work. The enclosures of the device are retrieved from the warehouse to be made available at the input of Work Cell 1. The operators at Op. 1-11 and Op. 1-21 use the drill presses to make several holes on the enclosure. At Op. 1-12 and Op.1-22, the operators affix labels on the exterior and interior of the enclosures. Then, at Op. 1-13 and Op. 1-23, the operators inspect the quality of the previous two steps to ensure locations of the holes and the labels comply to the product design. The intermediate products are then places on a conveyor to be transported to Work Cell 2. The operators at Op. 2-11 and Op. 2-21 install the printed circuit board, electrical wires, switch, and several small mechanical parts into the enclosure and fix them in place by applying screws into the holes drilled in Op. 1-11 and Op. 1-21. Finally, the operators at Op. 2-12 and Op. 2-22 inspect the assembled devices, test the electrical/mechanical functions, fix any issue discovered, and package the finished products with instruction manual cards into cardboard boxes.

B. Problem Description

As one can see, all operations in this production system involve human labor and the production issues/interruptions are often difficult to identify or trace. There is no production data recorded before, outside of the total number of units produced in a shift. On the other hand, the manufacturer did not have any in-house expertise and resources to solve this issue. To help the manufacturer enhance the system performance, the following problems are addressed:

- **Modeling**: Construct mathematical model for the production system based on the layout and collected data.
- **Analysis**: Use the mathematical model to quantitatively analyze the baseline system performance.
- **Improvement**: Based on the system model and the baseline analysis, investigate potential improvement actions to enhance system performance.

III. Mathematical Modeling

A. Structural and parametric modeling

Based on the system layout (Fig. 1), we build the complete model as Fig. 2, where the circles represent operations and the rectangle represents the in-process buffer connecting the two work cells. Note that no buffers are inside the work cells as the operators in the same work cells are expected to work synchronously. Then, we further aggregate the operations in each work cell to reach the simplified model.

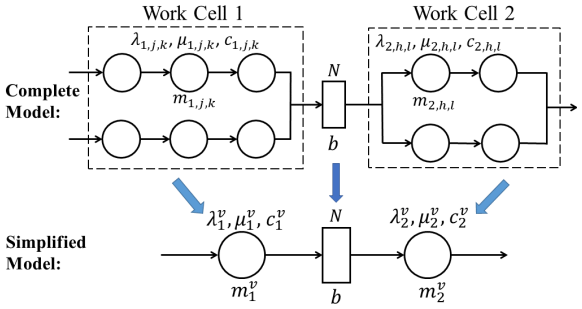

Fig. 2. Structural and parametric modeling

Since the operations may experience cycle overruns and delays, we model them as *unreliable machines* and assume the up- and downtime of a *machine* are exponential random variables with parameters λ (1/min) and μ (1/min), respectively. The parameters λ and μ are referred as the *breakdown rate* and *repair rate*, respectively. Given parameters λ and μ, *machine efficiency* e, is calculated as

$$e = \frac{\mu}{\lambda + \mu} = \frac{T_{up}}{T_{up} + T_{down}}, \tag{1}$$

where $T_{up} = 1/\lambda$ and $T_{down} = 1/\mu$ are the average up- and downtimes, respectively. In addition, the processing speed of an operation is denoted as c (parts/min), while the cycle time is denoted as τ (min). Thus, for the complete system model, an operation can be characterized by a vector: $(\lambda_{1,j,k}, \mu_{1,j,k}, c_{1,j,k})$, $j = 1, 2$ and $k = 1, 2, 3$, for the ones in Work Cell 1; and $(\lambda_{2,h,l}, \mu_{2,h,l}, c_{2,h,l})$, $h, l = 1, 2$, for the ones in Work Cell 2. Similarly for the simplified model, the aggregated/virtual machines are characterized by $(\lambda_i^v, \mu_i^v, c_i^v)$, $i = 1, 2$. Finally, the in-process buffer is characterized by its capacity N, i.e., the maximum number of parts that the buffer can hold (also see Fig. 2).

B. Model Parameter Identification

1) Data collected: In this system, we measure the buffer capacity based on the total length of the conveyor space and dimensions of the product, i.e., $N = 15$. The reliability parameters (λ and μ) of each operation were not available due to the challenges discussed in Section I. The average cycle time of each operation under typical conditions was measured as follows:

$$\boldsymbol{\tau_1} = \begin{bmatrix} 1.82 & 1.87 & 1.99 \\ 2.25 & 2.28 & 2.21 \end{bmatrix} \text{(min)}, \ \boldsymbol{\tau_2} = \begin{bmatrix} 1.56 & 1.65 \\ 1.83 & 1.76 \end{bmatrix} \text{(min)}. \tag{2}$$

Thus, the processing speeds of $m_{1,j,k}$'s and $m_{2,h,l}$'s in the ten-machine model can be obtained as follows:

$$\begin{aligned} \boldsymbol{c_1} &= \begin{bmatrix} 0.5482 & 0.5302 & 0.5025 \\ 0.4451 & 0.4386 & 0.4526 \end{bmatrix} \text{(parts/min)}, \\ \boldsymbol{c_2} &= \begin{bmatrix} 0.6379 & 0.6061 \\ 0.5464 & 0.5614 \end{bmatrix} \text{(parts/min)}. \end{aligned} \tag{3}$$

Using the parts flow data of the buffer measured during a 4-week span, the following performance metrics were obtained:

$$\begin{aligned} TP^* &= 0.7026 \text{ (parts/min)}, \qquad P_0^* = 0.2816 \\ WIP^* &= 5.6808 \text{ (parts)}, \qquad P_N^* = 0.0709, \end{aligned} \tag{4}$$

where these performance metrics are defined as follows:

- *Throughput*, TP: the average number of parts produced by machines $m_{2,1,2}$ and $m_{2,2,2}$ combined in the complete model or by machine m_2^v in the simplified model per time unit (e.g., minute, hour) during steady state;
- *Work-in-process*, WIP: the average number of parts contained in buffer b during steady state;
- *Probability that buffer b is empty*, P_0;
- *Probability that buffer b is full*, P_N.

2) Algorithm: According to the machine connection relationship presented in [3] and [14], the aggregated processing speeds of m_i^v can be calculated as:

$$\begin{aligned} c_1^v &= \sum_j \min\{c_{1,j,1}, c_{1,j,2}, c_{1,j,3}\} = 0.9411 \text{ (parts/min)}, \\ c_2^v &= \sum_h \min\{c_{2,h,1}, c_{2,h,2}\} = 1.1525 \text{ (parts/min)}. \end{aligned} \tag{5}$$

For a two-machine model, given c_i^v's and N, TP, WIP, P_0, and P_N are functions of λ_i^v's and μ_i^v's ($i = 1, 2$), which are derived in [3]. Instead of identifying $(\lambda_1^v, \mu_1^v, \lambda_2^v, \mu_2^v)$ directly, we identify the machine parameters in the form of $\mathbf{x} = (\mu_1^v, \mu_2^v, e_1^v, e_2^v)$, which is more convenient in expressing the feasible region of $\mathbf{x}$. Next, we define $\mathbf{F}(\mathbf{x})$ as:

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ f_3(\mathbf{x}) \\ f_4(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} TP(\mu_1^v, \mu_2^v, e_1^v, e_2^v) - TP^* \\ WIP(\mu_1^v, \mu_2^v, e_1^v, e_2^v)/N - WIP^*/N \\ P_0(\mu_1^v, \mu_2^v, e_1^v, e_2^v) - P_0^* \\ P_N(\mu_1^v, \mu_2^v, e_1^v, e_2^v) - P_N^* \end{bmatrix}. \tag{6}$$

where TP^*, WIP^*, P_0^*, and P_N^* are the observed system performance metrics given in (4). Based on the above, we formulate the following constrained optimization problem:

Find $\mathbf{x} = (\mu_1^v, \mu_2^v, e_1^v, e_2^v)$ *that minimizes the 2-norm of error function* $\mathbf{F}$ *over a certain box-constraint set* $\mathbf{X}$*, i.e.,*

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}) = ||\mathbf{F}(\mathbf{x})||^2, \\ &\textit{s.t. } \mathbf{x} \in \mathbf{X}, \end{aligned} \tag{7}$$

where $\mathbf{X} = \{\mathbf{x} \in \mathbb{R}^4 | \mu_l \le \mu_i^v \le \mu_u, e_l \le e_i^v \le e_u, i = 1, 2\}$ *and* μ_l*,* μ_u*,* e_l*, and* e_u *are the lower- and upper-bounds for the parameters* μ_i^v*'s and* e_i^v*'s.*

To solve problem (7), we propose a searching algorithm based on projected quasi-Newton method [15] in combination with the multi-start strategy of global search [16], called *multi-start modified projected quasi-Newton method* (MMPQN). We use $[\cdot]^+$ to denote the projection on the constraint set $\{\mathbf{x} \in \mathbb{R}^4 \mid a_i \le x_i \le b_i, i = 1, \ldots, 4\}$, and the ith coordinate of the projection of $\mathbf{x}$ is given by [17]:

$$[x_i]^+ = \begin{cases} a_i & \text{if} \quad x_i \le a_i, \\ b_i & \text{if} \quad x_i \ge b_i, \\ x_i & \text{otherwise.} \end{cases} \tag{8}$$

The projection arc is defined as $\bar{\mathbf{x}}(\alpha) = [\mathbf{x} - \alpha \mathbf{S} \nabla f(\mathbf{x})]^+$, where $\alpha \in (0, 1)$ and $\mathbf{S}$ is gradient scaling matrix. To find an appropriate α in each iteration, Armijo rule [17] is used and we find the smallest non-negative integer q such that

$$f(\mathbf{x}) - f(\mathbf{x}(\beta^q)) \ge \sigma \nabla f(\mathbf{x})^T [\mathbf{x} - \mathbf{x}(\beta^q)], \tag{9}$$

where $\beta, \sigma \in (0, 1)$. Then, $\alpha = \beta^q$. Next, to calculate $\mathbf{S}$ for next iteration after updating $\mathbf{x}$ along the feasible direction, we first divide the variables of $\bar{\mathbf{x}}$ into two groups: *free* and *restricted*. The latter refers to the subset of variables that are close to their bounds:

$$\begin{aligned} \mathfrak{R} = &\{i \mid \bar{x}_i < a_i + \epsilon, \ \partial_i f(\bar{\mathbf{x}}) > 0\} \cup \\ &\{i \mid \bar{x}_i > b_i - \epsilon, \ \partial_i f(\bar{\mathbf{x}}) < 0\}, \end{aligned} \tag{10}$$

where ϵ is very small positive number. The quasi-Newton-type gradient scaling only applies to the free variables, while the restricted variables descend along the gradient direction. In summary, we randomly select D initial points from $\mathbf{X}$ and the steps MMPQN are described as Algorithm 1.

Algorithm 1 Multi-start Modified Projected Quasi-Newton Method (MMPQN)

for $n = 1, \ldots, D$ **do**
  Initialization: Set
    $k = 0$, $\mathbf{x}^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, e_1^{(0)}, e_2^{(0)}) \in \mathbf{X}$ and
    $\mathbf{S}^{(0)} = \delta ||\nabla f(\mathbf{x}^{(0)})||^{-1} \mathbf{I}$, where $\delta \in (0, 1)$ [18].
  while $||\nabla f(\mathbf{x}^{(k)})|| \ge \epsilon_g$ **do**
    1. Find appropriate value for $\alpha^{(k)}$ using Armijo rule;
    2. Compute the projection arc:
      $\bar{\mathbf{x}}^{(k)} = [\mathbf{x}^{(k)} - \alpha^{(k)} \mathbf{S}^{(k)} \nabla f(\mathbf{x}^{(k)})]^+$;
    3. Compute the feasible direction: $d^{(k)} = \bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}$;
    4. Update $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + s \cdot d^{(k)}$, where $s \in (0, 1)$;
    5. Compute the approximated inverse of the Hessian matrix $\overline{\mathbf{S}}^{(k+1)}$ via *BFGS* method [18] ;
    6. Determine restricted variables set $\mathfrak{R}^{(k)}$ by (10);
    7. Set $\mathbf{S}^{(k+1)} = \mathbf{I}$, and then, let
      $\mathbf{S}_{i,j}^{(k+1)} = \overline{\mathbf{S}}_{i,j}^{(k+1)}$, $\forall i, j \notin \mathfrak{R}^{(k)}$;
    8. $k = k + 1$;
  end while
  $\widehat{\mathbf{x}}_n = \mathbf{x}^{(k)}$.
end for
Return $\widehat{\mathbf{x}} = \arg\min_{\widehat{\mathbf{x}}_n} f(\widehat{\mathbf{x}}_n)$.

Using the MMPQN method, we solve problem (7) with observed performance metrics (4) and obtain the parameters of the simplified two-machine model as follows:

$$\begin{aligned} T_{up,1}^v &= 12.68, \quad T_{down,1}^v = 3.03, \quad e_1^v = 0.8071, \\ T_{up,2}^v &= 19.04, \quad T_{down,2}^v = 7.66, \quad e_2^v = 0.7129. \end{aligned} \tag{11}$$

With the identified parameters (11), we estimate the performance metrics based on the two-machine model and evaluate the errors compared with the observed ones using

$$\epsilon_{TP} = \frac{|\widehat{TP} - TP^*|}{TP^*} \cdot 100\%, \qquad \epsilon_{P_0} = |\widehat{P}_0 - P_0^*|,$$
$$\epsilon_{WIP} = \frac{|\widehat{WIP} - WIP^*|}{N} \cdot 100\%, \quad \epsilon_{P_N} = |\widehat{P}_N - P_N^*|, \tag{12}$$

where $\widehat{\cdot}$ denotes the estimated performance metrics. As a result, both ϵ_{TP} and ϵ_{WIP} are below $10^{-6}\%$, while both ϵ_{P_0} and ϵ_{P_N} are below 10^{-6}. In other words, these identified parameters can provide an almost perfect match to observed performance metrics from the factory floor.

IV. System Baseline Analysis and Improvement

A. Blockage, starvation, and system bottleneck

With the identified model parameters, we can estimate those performance metrics which are difficult to be measured during production, such as blockage of Work Cell 1 (BL_1) and starvation of Work Cell 2 (ST_2). Using the two-machine line model, we obtain $BL_1 = 0.0605$ and $ST_2 = 0.1033$.

Furthermore, the system bottleneck (BN), which is the machine that leads to the maximal increase of TP when the processing speed of one machine is increased, can be identified. Using the bottleneck identification method introduced in [3], we obtain that the BN of the system is Work Cell 1, because $BL_1 < ST_2$. Moreover, we define the BN in Work Cell 1 as the machine that leads to the maximal increase of Work Cell 1's overall processing speed when the processing speed of one machine is increased. Based on (3) and (5), we can easily obtain that Op. 1-22 is the BN of Work Cell 1.

B. Effects of increasing machine processing speed

Since Work Cell 1 is the system BN, we consider increasing its processing speed to improve the system TP. With the identified parameters, we first find out the threshold point $\tilde{c}_1^v$ that makes the BN switch to Work Cell 2, i.e., $\tilde{c}_1^v = 1.1045$. We assume $\tilde{c}_1^v = 1.1045$ to be the upper bound of c_1^v improvement and the corresponding improved TP is 0.7735 part/min, which amounts to 10.09% improvement over the baseline throughput TP^*. Then, the increases of TP with $c_1^v \in [0.9411, 1.1045]$ are plotted in Fig. 3. Additionally, we set 2.5%, 5% and 7.5% as the TP improvement targets, namely, *minor*, *medium*, and *major* improvement, and then, we find the corresponding c_1^v which bring the TP to these improvement targets (see Fig. 3).

To achieve those particular c_1^v's, we create following improvement plans shown as Table I. Since the operators (usually trainees on the job) on the second parallel line have much lower processing speed than those on the first line, we prioritize their improvement, which can be accomplished through better training by more experienced operators. Besides, we improve the performance of the lower operators on the first line to let them gain the similar speed as the most proficient one. Finally, the manufacturer may explore the potential of designing and introduce new technologies to reach the maximum improvement.

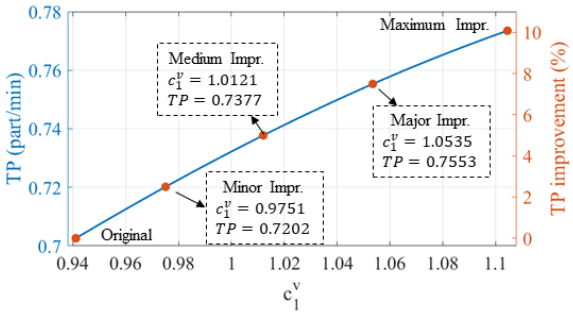

Fig. 3. TP and improvement (%) vs. c_1^v

TABLE I
Improvement Plans of Operation Speed-up in Work Cell 1

	$c_{1,j,k}$	c_1^v
Minor (2.5%)	$\begin{bmatrix} 0.5482 & 0.5302 & 0.5025 \\ \mathbf{0.4726} & \mathbf{0.4726} & \mathbf{0.4726} \end{bmatrix}$	0.9751
Improvement		
Medium (5%)	$\begin{bmatrix} 0.5482 & 0.5302 & \mathbf{0.5302} \\ \mathbf{0.4819} & \mathbf{0.4819} & \mathbf{0.4819} \end{bmatrix}$	1.0121
Improvement		
Major (7.5%)	$\begin{bmatrix} 0.5482 & \mathbf{0.5482} & \mathbf{0.5482} \\ \mathbf{0.5053} & \mathbf{0.5053} & \mathbf{0.5053} \end{bmatrix}$	1.0535
Improvement		
Max. (10.1%)	$\begin{bmatrix} \mathbf{0.5523} & \mathbf{0.5523} & \mathbf{0.5523} \\ \mathbf{0.5522} & \mathbf{0.5522} & \mathbf{0.5522} \end{bmatrix}$	1.1045
Improvement		

C. Effects of reducing machine downtime

Reducing downtime of the operations is also a commonly used method to improve system throughput. In this case study, since the up- and downtime data of each individual operation are not available, we first estimate the individual machines' parameters ($T_{down,1,j,k}$, $e_{1,j,k}$, $T_{down,2,h,l}$ and $e_{2,h,l}$) based on the identified parameters of the aggregated machines ($T_{down,1}^v$, e_1^v and $T_{down,2}^v$ and e_2^v). In practice, the operations in the same work cell typically have similar performance through work balancing efforts. Therefore, we assume that each parallel line in a work cell has identical aggregated downtime and stand-alone throughput and also, the consecutive machines in the same line has identical downtime and stand-alone throughput. Based on the formulas to approximate the aggregated parameters of a group of parallel machines or consecutive machines derived in [14], we reversely calculate the individual machines' parameters from those of the aggregated machines. Generally, supposing Work Cell i consists of W parallel lines and each line contains Z consecutive machines, the algorithm is as follows.

1) Let $T_{down,i,j}^{par}$, $e_{i,j}^{par}$ and $c_{i,j}^{par}$ denote the *estimated* average downtime, efficiency and processing speed of the jth parallel line, and then we have

$$c_{i,j}^{par} = \min\{c_{i,j,1}, ..., c_{i,j,Z}\}, \tag{13}$$

$$e_{i,j}^{par} = \frac{c_i^v e_i^v}{W c_{i,j}^{par}}, \qquad T_{down,i,j}^{par} = W T_{down,i}^v. \tag{14}$$

2) Let $T_{down,i,j,k}^{con}$ and $e_{i,j,k}^{con}$ denote the *estimated* average downtime and efficiency of the kth machine on the jth line, and then we have

$$e_{i,j,k}^{con} = \frac{(e_{i,j}^{par} \prod_{k=1}^{Z} c_{i,j,k})^{\frac{1}{Z}}}{c_{i,j,k}}, \tag{15}$$

$$T^{con}_{down,i,j,k} = \frac{T^{par}_{down,i,j} \sum_{k=1}^{Z}(1 - e^{con}_{i,j,k})}{1 - e^{par}_{i,j}} \tag{16}$$

With the above algorithm, we obtain the estimated parameters (denoted as $\hat{\cdot}$) of each individual machine as follows:

$$\hat{\boldsymbol{T}}_{\boldsymbol{down,1}} = \begin{bmatrix} 6.59 & 6.59 & 6.59 \\ 6.34 & 6.34 & 6.34 \end{bmatrix}, \hat{\boldsymbol{T}}_{\boldsymbol{down,2}} = \begin{bmatrix} 16.79 & 16.79 \\ 16.40 & 16.40 \end{bmatrix};$$
$$\hat{\boldsymbol{e}}_{\boldsymbol{1}} = \begin{bmatrix} 0.8750 & 0.9048 & 0.9546 \\ 0.9538 & 0.9679 & 0.9380 \end{bmatrix}, \hat{\boldsymbol{e}}_{\boldsymbol{2}} = \begin{bmatrix} 0.8025 & 0.8446 \\ 0.8834 & 0.8511 \end{bmatrix}. \tag{17}$$

Due to the parallel structure in both work cells, we reduce the downtime of the operations carrying out the same processing work simultaneously (e.g., $T_{down,1,1,1}$ and $T_{down,1,2,1}$ together) to mimic the alleviation of certain common causes for failures/stoppages. Given the value of reduction factor a, the modified individual machine downtime, $(1-a\%)T_{down,1,j,k}$ or $(1-a\%)T_{down,2,h,l}$, is first used to update its own efficiency, $e_{1,j,k}$ or $e_{2,h,l}$. Then, with the modified downtime and efficiency, the aggregated machine parameters, e^v_i and $T^v_{down,i}$, of the two-machine model can be computed via the algorithm in [14]. Finally, the improved TP is calculated based on the two-machine model. The resulting improvement in TP are shown in Fig. 4, assuming the maximum of reduction factor of each T_{down} is 50%.

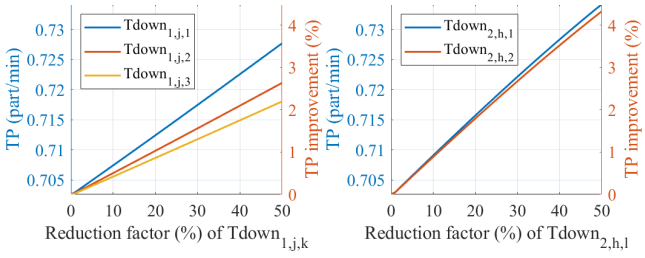

Fig. 4. TP and improvement (%) vs. T_{down} reduction

Based on Fig. 4, with the same reduction factor, reducing the T_{down} of Work Cell 2 leads to larger TP improvement. In particular, when we reach the maximum reduction on the T_{down} of Op. 2-11 and Op. 2-21, the largest possible TP can be obtained as $TP = 0.7342$ (part/min), which is about 4.50% higher than the baseline TP^*. Thus, only minor improvement target (2.5%) of TP can be achieved via reducing T_{down} and the lowest necessary reductions of $T_{down,1,j,1}$, $T_{down,1,j,2}$, $T_{down,2,h,1}$, $T_{down,2,h,2}$, $(j,h = 1,2)$ are 36%, 48%, 29% and 30%, respectively.

D. Effects of increasing buffer capacity

Besides operation speed-up and downtime reduction, the system throughput can be improved via buffer expansion. Specifically, with the identified model parameters (11), we increase the buffer capacity from $N = 15$ to its double size $N = 30$ and estimate the throughput of the system. The resulting effects of buffer capacity increase are illustrated in Figure 5. As shown in this figure, buffer expansion leads to 4.40% improvement on system throughput when the buffer capacity is doubled. To reach the minor improvement target, the lowest needed buffer capacity is 22.

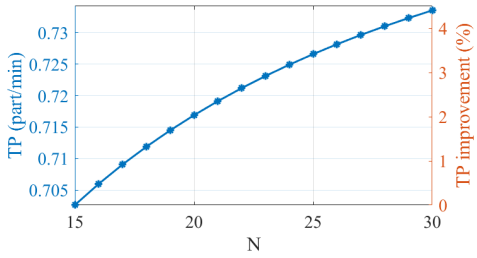

Fig. 5. TP and improvement (%) vs. N

These recommendations were submitted to the plant for consideration of implementation.

V. Model Sensitivity Analysis

Although the parameters identified for the simplified two-machine model can perfectly fit the observed system performance metrics, it is important to show whether this model is still robust when some model parameters (i.e., c, N, etc.) are changed for improvement. Furthermore, with the *identical downtime and identical stand-alone throughput* assumptions applied to the ten-machine model in Subsection IV-C, the identified parameters may be far from the actual up- and downtime of the operations. This may lead to deviation of the performance metrics from the true values when we design improvement actions. Thus, in this section, we study the robustness and sensitivity of the model when the "true" system parameters are different from the identified ones.

To carry this out, we first search for a set of machine parameters that leads to the same performance metrics of (4) under the ten-machine model. A solution is given below:

$$\boldsymbol{T}_{\boldsymbol{down,1}} = \begin{bmatrix} 34.85 & 30.62 & 32.25 \\ 31.11 & 32.84 & 29.67 \end{bmatrix}, \boldsymbol{T}_{\boldsymbol{down,2}} = \begin{bmatrix} 7.05 & 5.92 \\ 7.15 & 6.26 \end{bmatrix};$$
$$\boldsymbol{e}_{\boldsymbol{1}} = \begin{bmatrix} 0.9212 & 0.9153 & 0.9358 \\ 0.9020 & 0.9212 & 0.9385 \end{bmatrix}, \boldsymbol{e}_{\boldsymbol{2}} = \begin{bmatrix} 0.8372 & 0.8635 \\ 0.8581 & 0.8465 \end{bmatrix}. \tag{18}$$

As one can see, these parameters are quite different from the ones calculated in (17). In these experiments, we vary the processing speed, operation downtime and buffer capacity to mimic the potential improvement scenarios (similar to Section IV). With the new c's, T_{down}'s or N's, the estimated performance metrics are computed based on the two-machine line model and compared with the performance metrics of the modified system evaluated using simulation.

Similarly to Subsection IV-B, c^v_1 is selected to be modified. We uniformly select $c^v_1 \in [0.9411, 1.1045]$ as $c^{v,imp}_1$ and assign each $c_{1,j,k}$ based on following policy. If $c^{v,imp}_1/2 < \min\{c_{1,1,k}\}$, then let $\min\{c_{1,2,k}\} = c^{v,imp}_1 - \min\{c_{1,1,k}\}$, and all $c_{1,1,k}$'s are fixed. Otherwise, let $\min\{c_{1,j,k}\} = c^{v,imp}_1/2, \forall j$. With these $c_{1,j,k}$'s, fixed $c_{2,h,l}$'s and the "true" parameters (18), we compute the "true" performance metrics of the system by simulation. The estimation errors of TP and WIP obtained from the two-machine model are summarized in Fig. 6. Clearly, the model identified using the proposed approach still maintains high accuracy in estimating TP and WIP, even through the the machine parameters are different and $c_{1,j,k}$'s are modified.

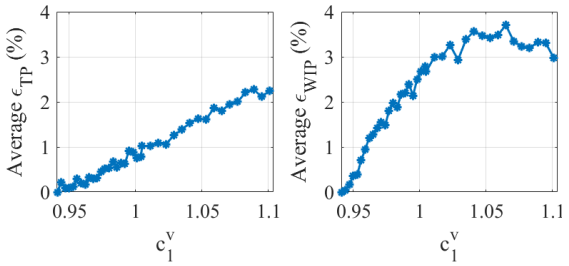

Fig. 6. TP and WIP error (%) vs. c_1^v

Then, with the "true" parameters (18), we calculate system TP and WIP using simulations for the downtime reduction cases at each operation. The average errors, when compared to those calculated in Subsection IV-C, are illustrated in Fig. 7. Despite the increasing trend of the errors when the reduction factors rise, the system performance estimates still remain in a relatively reasonable range.

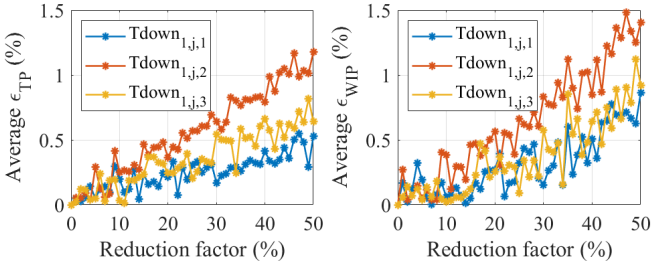

Fig. 7. TP and WIP error (%) for reducing $T_{down,1,j,k}$

Similarly, We calculate the performance metrics with the "true" parameters (18) using simulations for different buffer capacity up to $N = 30$. The errors of TP and WIP calculated from the two-machine model are plotted in Fig. 8. The errors of both performance metrics are low for all cases studied: below 2% for TP and 3% for WIP.

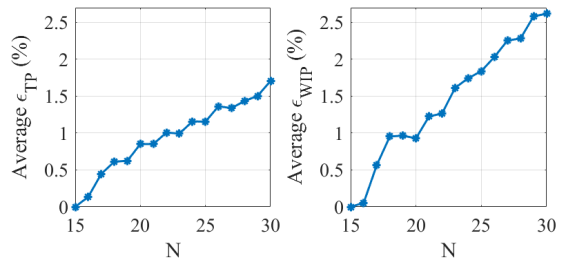

Fig. 8. TP and WIP error (%) vs. N

As demonstrated by the model sensitivity analysis, our approach can predict the performance metrics quite accurately not only under nominal parameters but also in a reasonable range of parameter change, all without any knowledge of T_{up} and T_{down} of individual machines. Such property is critical in designing improvement projects in manufacturing practice.

VI. Conclusion

In this paper, we apply a novel approach to modeling a production system in a case study at a local small manufacturer of medical devices. Specifically, we transform the complete ten-machine system model into a simplified two-machine asynchronous exponential line model. The model parameters are identified using multi-start modified projected quasi-Newton method to match the given performance metrics derived from the collected parts flow data. With the identified model parameters, we analyze the baseline performance and develop improvement actions to enhance its productivity. Finally, we study the sensitivity of the identified model through extensive numerical experiments. Future work includes extending this new modeling approach to more complex production system models, such as multi-stage production line models and assembly system models, and generalize the method to systems with machines following other reliability models (e.g. geometric, Gamma).

References

[1] H. T. Papadopoulos, C. Heavy, and J. Browne, *Queueing Theory in Manufacturing Systems Analysis and Design*. Chapman & Hill, London, UK, 1993.

[2] S. B. Gershwin, *Manufacturing Systems Engineering*. Prentice Hall, Englewood Cliff, NJ, 1994.

[3] J. Li and S. M. Meerkov, *Production Systems Engineering*. Springer, 2009.

[4] G. Liberopoulos and P. Tsarouhas, "Reliability analysis of an automated pizza production line," *Journal of Food Engineering*, vol. 69, no. 1, pp. 79–96, 2005.

[5] K. Park and J. Li, "Improving productivity of a multi-product machining line at a motorcycle manufacturing plant," *International Journal of Production Research*, vol. 57, no. 2, pp. 470–487, 2019.

[6] S.-Y. Chiang, "Bernoulli serial production lines with quality control devices: Theory and application," *Mathematical Problems in Engineering*, vol. 2006, 2006.

[7] Y. Feng, X. Zhong, J. Li, and W. Fan, "Analysis of closed-loop production lines with bernoulli reliability machines: Theory and application," *IISE Transactions*, vol. 50, no. 3, pp. 143–160, 2018.

[8] S. H. Huang, J. P. Dismukes, J. Shi, Q. Su, G. Wang, M. A. Razzak, and D. E. Robinson, "Manufacturing system modeling for productivity improvement," *Journal of manufacturing systems*, vol. 21, no. 4, pp. 249–259, 2002.

[9] Y. Sun, T. Zhu, L. Zhang, and P. Denno, "Parameter identification for bernoulli serial production line model," *IEEE Transactions on Automation Science and Engineering*, 2020.

[10] Y. Sun and L. Zhang, "Parameter identification for multiple-machine Bernoulli lines using statistical learning methods," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2020.

[11] ——, "Parameter identification for synchronous two-machine exponential production line model," in *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2021, pp. 1884–1889.

[12] T. A. Reddy and K. K. Andersen, "An evaluation of classical steady-state off-line linear parameter estimation methods applied to chiller performance data," *Hvac&R Research*, vol. 8, no. 1, pp. 101–124, 2002.

[13] R. Anish and K. Shankar, "Identification of nonlinear structural parameters using combined power flow and acceleration matching approaches," in *Advances in Mechanical Engineering*. Springer, 2020, pp. 1139–1149.

[14] C.-B. Yan and Q. Zhao, "A unified effective method for aggregating multi-machine stages in production systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 7, pp. 1674–1687, 2013.

[15] M. Schmidt, D. Kim, and S. Sra, "Projected newton-type methods in machine learning," *Optimization for Machine Learning*, no. 1, 2012.

[16] D. Peri and F. Tinti, "A multistart gradient-based algorithm with surrogate model for global optimization," *Communications in Applied and Industrial Mathematics*, vol. 3, no. 1, pp. 1–22, 2012.

[17] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, Belmont, MA, 2016.

[18] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.