Standalone closed-form formula for the throughput rate of asynchronous normally distributed serial flow lines

Adam Aboutaleb*, Parminder S. Kang, Raouf Hamzaoui, Alistair Duffy

School of Engineering and Sustainable Development, De Montfort University, Leicester, United Kingdom

Email: aaboutaleb@live.com; kang.parminder@gmail.com; rhamzaoui@dmu.ac.uk; apd@dmu.ac.uk

*Corresponding Author:

Adam Aboutaleb

School of Engineering, Qi Liping Campus, Shaoyang University, 207 National Road, Daxiang, Shaoyang, Hunan, China

Tel.: +86 182 7439 4443

Email: a.aboutaleb@live.com

Standalone closed-form formula for the throughput rate of asynchronous normally distributed serial flow lines

Flexible flow lines use flexible entities to generate multiple product variants using the same serial routing. Evaluative analytical models for the throughput rate of asynchronous serial flow lines were mainly developed for the Markovian case where processing times, arrival rates, failure rates and setup times follow deterministic, exponential or phase-type distributions. Models for non-Markovian processes are non-standalone and were obtained by extending the exponential case. This limits the suitability of existing models for real-world human-dependent flow lines, which are typically represented by a normal distribution. We exploit data mining and simulation modelling to derive a standalone closed-form formula for the throughput rate of normally distributed asynchronous human-dependent serial flow lines. Our formula gave steady results that are more accurate than those obtained with existing models across a wide range of discrete data sets.

Keywords: Serial flow lines, Flexible manufacturing systems, Throughput rate, Non-exponential stochastic processes, Data mining.

1. Introduction

During the past decades, several manufacturing systems were developed to keep pace with the advancements in technology and tailor products to customer needs. Customer requirements tend to be trending upwards in terms of complexity, which requires reshaping the manufacturing process to be flexible to handle complex products [1]. Flexible flow lines are an example of manufacturing systems that use flexible processes. Flexible flow lines are a cost-effective solution that combines the benefits of mass production and mass customisation strategies [2]. Such flow lines standardise the serial routing for all product variants while allowing manufacturing flexibility to take place at the process level to adapt to the product complexity. Flexible human-dependent processes, such as in the construction industry, can produce a range of products with variable complexity. They are less affected by setups and failures but have more stochastic processing times [3] due to the flexibility of the human brain, cognitive functions, skills and emotions [4]. However, with the increased flexibility and the resulting variability at the process level (Figure 1), production and process planning to maintain the performance targets becomes a challenging task.

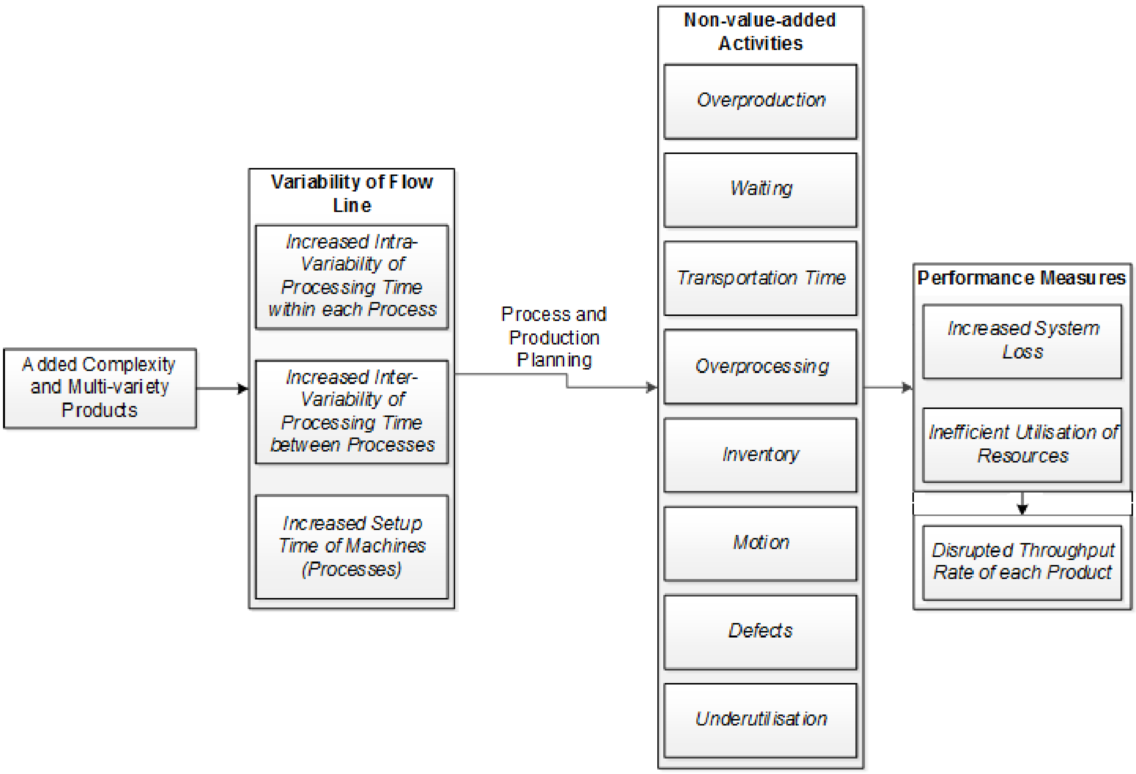

Figure 1. Impact of product complexity on manufacturing systems.

Several mathematical, simulation and empirical models were developed for different types of flow lines to assist production and process planners by estimating the effect of variability on the performance [1,5–9]. Research in the area of evaluative modelling focused on machine-based Markovian flow lines which are widely used in the manufacturing industry. Such Markovian models were developed primarily to include queue capacity and repair and failure rates. However, as they only assume deterministic, exponential or phase based distribution of the processing times, they are not applicable for human-dependent processes. For stochastic non-Markovian processes that follow a distribution other than the exponential or phase-type distribution, analytical methods do not exist [10] and simulation and empirical approaches are used instead.

Modern simulation modelling software provides a visual platform with high flexibility to accurately represent complex flow lines [11]. However, simulations are usually case-based and time-consuming. Closed-form formulas can be generic, simple, time efficient and relationships are easily understood [3,6,7,12]. Data mining of simulated data has been the main route for the empirical approach [7]. While empirical formulas are not mathematically proven to be correct, they can provide a reliable model to estimate the throughput rate and optimise the planning and operations of flow lines.

Standalone closed-form formula of the throughput rate

To the best of our knowledge, no standalone closed-form empirical formula exists for the throughput rate of asynchronous flow lines with normally distributed process variability. Our paper fills this gap by

- proposing a generic representation of arbitrary length human-dependent non-exponential flow lines using nonlinear terms. This allowed for an accurate closed-form modelling of the throughput rate; and
- developing a standalone closed-form empirical formula to estimate the throughput rate of asynchronous flow lines with normally distributed process variability to a higher accuracy than existing models. The validity of the proposed formula to real-world scenarios was successfully tested through a wide range of representative data sets.

Section 2 discusses related work. Section 3 gives a simple empirical standalone closed-form formula for the throughput rate of asynchronous human-dependent serial flow lines. The formula is tested and analysed in Section 4 and validated in a real-world industrial case study in Section 5. Finally, Section 6 concludes the paper and suggests future work.

2. Related Work

2.1 Variability in Human-dependent Flexible Flow Lines

Flexibility in manufacturing systems is a measure of the capability of processes to adapt and the control system to take a different decision in response to changes within the manufacturing system [13–15]. Sethi and Sethi [16] identified three levels of flexibility: component, system and aggregated. Wiendahl et al. [17] identified three perspectives to classify manufacturing flexibility: order, product and resource. Windt and Jeken [18] combined the two concepts and added another sub-category, allocation flexibility (Figure 2). Accordingly, flexible manufacturing systems generate multiple degrees of variability which will eventually transfer to the performance targets. Variability can occur in flexible flow lines due to:

i. sudden interruptions to the flow line as a result of failures or setup time when a product is replaced by another one;
ii. restricted queue capacity of work-in-progress [19];
iii. production of customisable products according to customers' demand, where a single flow line produces different options and features of a product [20];

iv. constraints and differences between related products and their associated processes [21];

v. priority of the customer orders, e.g., rush orders [22];

vi. homogeneity of the flow line: variability of processing time from one process to another along the flow line can be zero, i.e., homogenous, or changing, i.e., inhomogeneous [21]; and

vii. natural reasons, such as the fluctuations in human's cognitive functions and emotions [4,12].

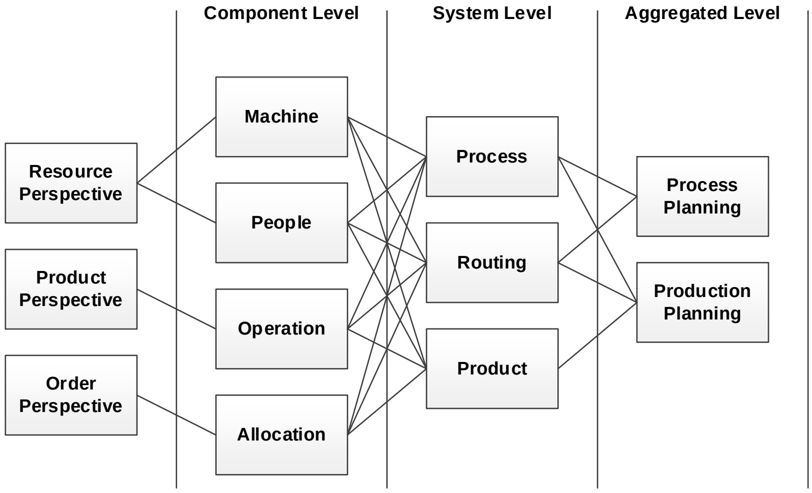

Figure 2. Classification of manufacturing flexibility (based on [18]).

This paper focuses on human-dependent processes. Hence, machine reliability, setup time and queue capacity are not of a concern, and variability is assumed to be primarily due to the intra- and inter-variability of processing times (see iii to vii above).

The normal distribution is known to be the most applicable form of distribution pattern that represents the variability of human-dependent activities [12,23]. For non-exponential flow lines, including normally distributed ones, several empirical studies [3,6,10,24–26] suggest that the variability corresponds to a coefficient of variation (ratio of standard deviation and mean) $c \leq 1$.

A limitation of the normal distribution in this context is that its range is $(-\infty, \infty)$ while processing times are non-negative. A solution to this problem is to accurately estimate the coefficient of variation, change the support to $(0, \infty)$ and enforce the probability density function to be zero when the processing time is negative.

2.2 Current Evaluative Models of Flow Lines

For short Markovian flow lines, both exact mathematical models [1,5,27–29] and closed-form formulas [3,12,26,30,31] exist.

Recent work [6–9,32,33] used approximate analytical solutions, such as decomposition and aggregation methods, to model arbitrary length Markovian flow lines.
However, Markovian flow lines are not suitable for human-dependent processes which are known [12,23] to have normally distributed variability patterns.
Li and Meerkov [6] proposed the following formula for the throughput rate (TR) of non-Markovian asynchronous non-exponential serial flow lines consisting of N processes $P_i, i = 1,2,...., N$ with coefficient of variation $c_i, i = 1,2,...., N$:

$$TR = TR^d - \left(TR^d - TR^e\right)c_{av} \text{ where } TR^d = \frac{1}{\mu_{\max}} \text{ and } c_{av} = \frac{1}{2N}\sum_{i=1}^{N} c_i \tag{1}$$

Here TR^d is the throughput rate of the serial flow line if it is assumed that the N processes have deterministic processing time (thus $c_i = 0, i = 1,2,...., N$) and TR^e is the throughput rate of the serial flow line if it is assumed that the N processes have exponential processing time (thus $c_i = 1, i = 1,2,...., N$).

However, the variable TR^e needs to be obtained using simulations so the formula cannot be applied on its own for performance analysis. Recent work [3,30] presents an interesting Markov chain-based analytical model to obtain TR^e for short service-based flow lines with non-exponential processing times.

3. Method

3.1 Notations

A list of symbols used in this paper is given in Appendix A.

3.2 Assumptions

We focus on flexible flow lines with a standard serial flow line arrangement, infinite queue capacity and stochastic human-dependent processes that follow the rules of normal distribution (Figure 3). The infinite queue capacity assumption here means that the processes cannot stop due to full buffer capacity, i.e., the process has two states; either processing the work item or waiting for work-in-progress to arrive.

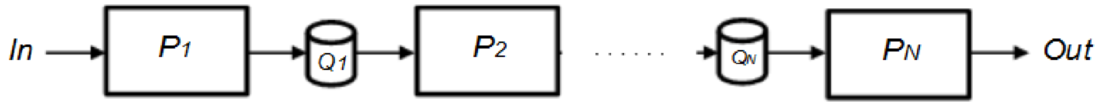

Figure 3. Serial flow line.

Hence, the following assumptions are made:

1. The flow line consists of serial processes.

2. The processing time for each process P_i is independent of the upstream and downstream processes P_{i-1} and P_{i+1}, i.e., the flow line is asynchronous.
3. The time for each process P_i is normally distributed with mean processing time $\mu_i, i = 1,2,...., N$ and standard deviation $\sigma_i, i = 1,2,...., N$.
4. The process P_i is reliable.
5. Blocking of a process P_i can only occur when it completed processing a part while the downstream process P_{i+1} is still busy and no queue exists between them.
6. A process P_i can get '*starved*' when the upstream process P_{i-1} is not completed.
7. Required resources (machine, people, tools, etc.) are always available at the respective process P_i.
8. If the process P_i is not '*blocked*' or '*starved*', it is in a '*busy*' state, i.e., the process P_i is not allowed to be '*idle*'.
9. The flow line is saturated, i.e., the first process P_1 is never '*starved*' for inputs, e.g., materials, orders, and the last process P_N is never '*blocked*', i.e., it has infinite queue capacity.
10. The travel time between processes is zero, i.e., transportation of materials and work in progress is modelled as a separate process.
11. The loss rate in throughput rate TR is zero, i.e., there are no defective products.

3.3 Design

From the model building perspective, the three main process-based parameters that represent the flow line, μ_i, c_i and N , do not remain constant for asynchronous flow lines. Hence, a set of generic parameters were studied to represent variability within the flow line with minimal number of variables.

The maximum processing time plays an important role in the throughput rate of non-exponential flow lines. In fact, the bottleneck, i.e., the process with the maximum actual processing time, governs the throughput rate for deterministic processing times (first term in Equation 1). However, for non-deterministic flow lines (second term in Equation 1), such as the case in this paper, the process with the maximum mean and maximum actual processing time do not always match.

The bottleneck can constantly move based on the mean processing times along the flow line and the average coefficient of variation, i.e., when the actual processing time of a process exceeds the maximum mean processing time.

Furthermore, the use of processing times of each process, in addition to what was explained in Section 2, will require an enormous number of independent variables to represent long flow lines.

Hence, an additional generic data mining-compatible parameter which is the average of mean processing times within the flow line can explain the discrepancy due to the potential movement of the bottleneck for asynchronous non-exponential flow lines. It essentially represents the proximity of the processing times with respect to the flow line, hence, the potential movement of the bottleneck.

Furthermore, researchers did not investigate the effect of the location of the process with the maximum mean processing time within the flow line, i.e., the ratio between the process with the maximum mean processing time and the length of the flow line.

Hence, these two parameters were added along with the ones from Li and Meerkov's formula (Equation 1) as follows:

- Maximum Mean Processing Time within Flow Line $(\mu_{\max})$:

$$\mu_{\max} = \max_i \mu_i \tag{2}$$

- Average Coefficient of Variation (c_{av}):

$$c_{av} = \frac{1}{2N}\sum_{i=1}^{N} c_i \tag{3}$$

- Length (N)
- Average Mean Processing Time within Flow Line (μ):

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \mu_i \tag{4}$$

- Location Ratio of the Process with Maximum Mean Processing Time (l):

$$l = {}^{i}\!/_{N} \text{ such that } \mu_i = \mu_{\max} \tag{5}$$

Furthermore, the investigation included the direct and multiplicative inverse of linear and nonlinear terms of each variability parameter, i.e., variable. The general criteria for election of parameter terms as model predictors were set as:

i. Only terms with highly strong relationships to TR were considered, i.e., correlation coefficient equals or higher than 0.8;

ii. Relationship is considered insignificant and the predictor terms excluded if the *p*-value is higher than 0.1 with the following levels [34] used for evaluation of the significance:
    a. Highly significant: *p*-value is less than 0.01;
    b. Statistically significant: *p*-value is higher than 0.01 but less than 0.05;
    c. Possibly significant: *p*-value falls between 0.05 and 0.1; and
    d. Insignificant: *p*-value is higher than 0.1.

The regression covariates in stepwise regression were also selected using these criteria.

3.4 Data Mining Methodological Framework

The data mining framework is based on a search approach that investigates the degrees of freedom (DOFs) at each of the three phases of the model development phases: data pre-processing, feature selection and model building.

Figure 4 illustrates the methodological framework. Synthetic data were generated and complexity was introduced gradually to the data set to cover a wide range of variability scenarios that can occur in asynchronous human-dependent serial flow line. The variability parameters were extracted from the datasets and simulations were applied to obtain the simulated throughput rate for each variability scenario within the datasets. Statistical analysis was then applied to shortlist the variability parameters based on their impact on the simulated throughput rate. Finally, the throughput rate model was built, using supervised machine learning techniques, as a function of the shortlisted variability parameters and validated against the simulated throughput rate. The developed model was then validated with continuous actual data from a real-world industrial case study.

3.4.1 Data Pre-processing

Synthetic data were sampled into two classes; Class I (the training set), and Class II (the test set). The steady state throughput rate was obtained using simulations with a confidence interval of 95%.

3.4.2 Feature Selection

This phase provides a new representation of asynchronous non-exponential serial flow lines using selected linear and nonlinear terms of line-based parameters based on their impact on TR. The investigation included statistical impact and stability analysis of each prediction line-based parameter on TR.

A data set (Class III) was created with the smallest number of changes for each parameter. Each data set included a number of sub-sets to verify the results (Table 1).

Table 1. Details of data set – Class III for the first four parameters.

Parameter	No of Sub-sets	Range
$\mu_{\max}$	2	$\mu_{\max} \in \{2,3,...,10\}$
l	3	$l \in \{1,2,...,15\}$
c_{av}	3	$c_{av} \in \{0,0.01,0.025,0.05,0.075,0.1,0.25,0.5,0.75,1\}$
N	4	$N \in \{4,6,8,10,12,13,15,17,19,21,23,25,27,29\}$

As for μ, eight sub-sets were created. c_{av} was kept high at 0.75 for all sub-sets to allow the bottleneck to move from one process to another. All sub-sets have a wide and fixed range of maximum mean processing time, $\mu_{\max} \in \{2,3,...,10\}$, and $N = 15$.

The subsets were arranged based on the proximity of μ and $\mu_{\max}$ from least to highest as follows:

$$III - \mu - 1 : \mu_i = \begin{cases} 1 & for\ N \in \{1,2,...,7\}, \\ \mu_{\max} & for\ N = 8, \\ 1 & for\ N \in \{9,10,...,15\}, \end{cases} \qquad III - \mu - 2 : \mu_i = \begin{cases} \mu_{\max} & for\ i < 0.5N, \\ 1 & for\ i > 0.5N, \end{cases}$$

$$III - \mu - 3 : \mu_i = \begin{cases} 1 & for\ i \le 0.538N, \\ \mu_{\max} & for\ i > 0.538N, \end{cases} \qquad III - \mu - 4 : \mu_i = \begin{cases} \mu_{\max} & for\ N \in \{1,2,...,7\}, \\ 1 & for\ N = 8, \\ \mu_{\max} & for\ N \in \{9,10,...,15\}, \end{cases}$$

$$III - \mu - 5 : \mu_i = \begin{cases} \mu_{\min} & for\ N \in \{1,2,...,7\}, \\ 60 & for\ N = 8, \\ \mu_{\min} & for\ N \in \{9,10,...,15\}, \end{cases} \qquad III - \mu - 6 : \mu_i = \begin{cases} 60 & for\ i < 0.5N, \\ \mu_{\min} & for\ i > 0.5N, \end{cases}$$

$$III - \mu - 7 : \mu_i = \begin{cases} \mu_{\min} & for\ i \le 0.538N, \\ 60 & for\ i > 0.538N, \end{cases} \qquad III - \mu - 8 : \mu_i = \begin{cases} 60 & for\ N \in \{1,2,...,7\}, \\ \mu_{\min} & for\ N = 8, \\ 60 & for\ N \in \{9,10,...,15\} \end{cases}$$

Statistical analysis was carried out to determine the strength and significance of the relationship between parameters, including its linear and nonlinear terms, and TR. Correlation analysis was applied to examine the strength of the relationship. However, to determine the significance of this relationship, Analysis of Variance (ANOVA) was performed on the data set; *f*- and *p*-value of regression coefficients and *f*-value of regression model were examined to determine if the parameter is statistically significant. Finally, best sub-set regression was applied to verify the results and determine if a parameter can be excluded from the model building stage.

Best sub-sets regression is a method that can be used to do this validation in one step since it will provide statistical measures for the best single-variable model, 2-variables, etc.

3.4.3 Model Building

An evaluative model for asynchronous flow lines was built during this phase. The model used different cross validation partitions to suit the nature of data sets and for comparison purposes. The DOFs in this phase are:

- **Supporting Predictors:** selection of the supporting predictors to be included in the training of the data mining model;
- **Cross Validation Partitioning:** data sets used as training and test sets; and
- **Modelling Method:** supervised machine learning regression to build the formula-based evaluative model of the throughput rate:
  a. *Stepwise Regression – Model Type*:
     - *Interaction of linear terms, i.e., covariates can be a single or multiplication of two linear predictor term(s);*
     - *Pure quadratic;*
     - *Quadratic;*
     - *Polynomial up to the 6th degree;*
  b. *Stepwise Regression – Bounded and Unbounded Steps*;
  c. *Robust Regression:* eight fitting techniques for the least squares; and
  d. *Regularisation Algorithms:* three algorithms for regularisation of the least squares.

The model building process ran through the different degrees of freedom for the data sets representing flow lines. It started by importing all the data sets $\mathbf{D}$. The individual data set $I/II-A-1$ to $I/II-A-8$ were segregated to $\mathbf{S}_n, n \in \{1,2,...,8\}$. Multiple training data sets of $\mathbf{S}_n$ can be included within the data set pool $\mathbf{D}_x, x \in \{1,2,..., X\}$, where the variable X determines the number of data sets within $\mathbf{D}_x$ that can be used for training when n increases, after exclusion of the best performing data set $\mathbf{S}_n$ from the data set pool $\mathbf{D}_x$, $X = 8 - n + 1$. The process-based parameters, as selected in Phase II, were then identified as the model main predictors $p_y, y \in \{1,2,3,4,5\}$ and supporting predictors $p_j, j \in \{1,2,3,4\}$ such as the predictors set $\mathbf{P} \subset \mathbf{D}$.

Standalone closed-form formula of the throughput rate

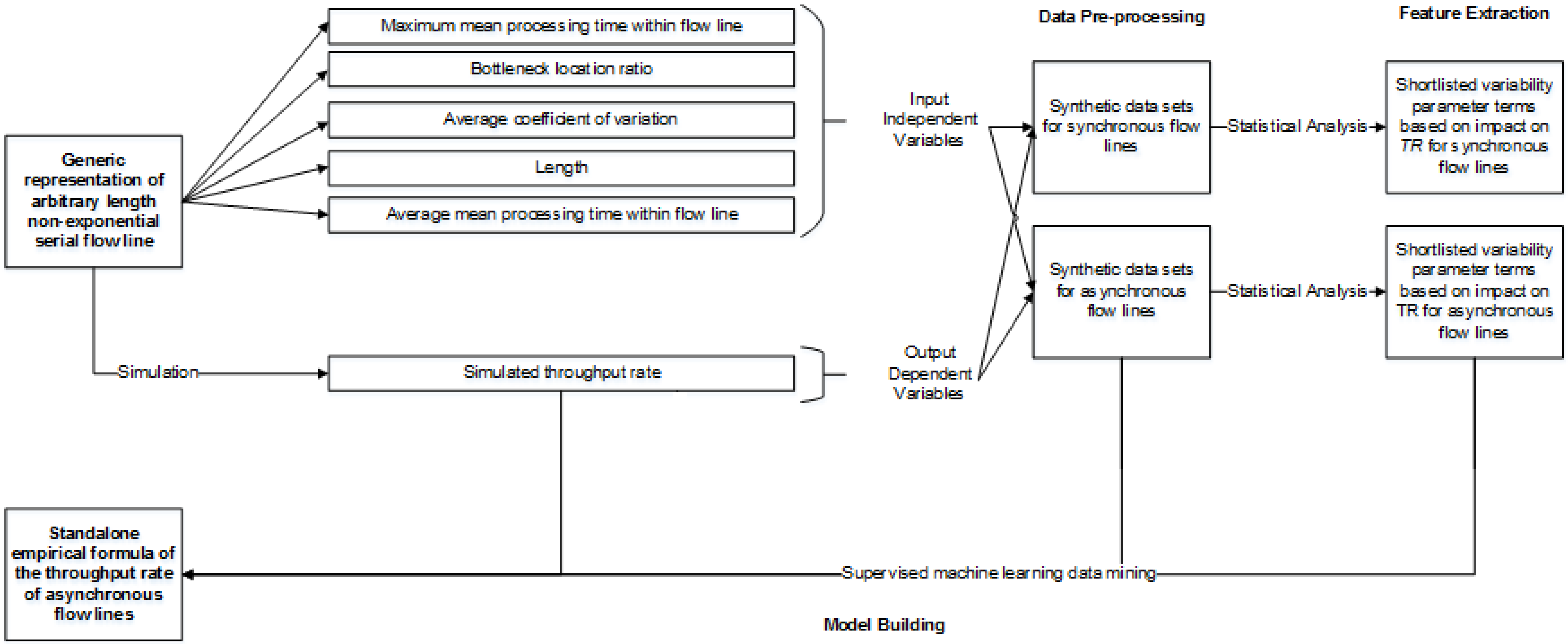

Figure 4. Data mining methodological framework.

A counter of the training set number w was then started. The models were trained and tested using the cross validation technique with step partitioning of the training and test sets, i.e., iterative selection from the data set $\mathbf{D}$. The data sets used for training were excluded from the main test set $\mathbf{T}_o$ but included in the supporting test set $\mathbf{T}_u$ to examine the models for overfitting.

Each method m was run and the mean absolute percentage error for each test set $\mathbf{T}_a$ was calculated as follows:

$$e_{ma} = \frac{100}{q_{T_a}} \sum \left| \frac{\mathbf{TR}_a^{sim} - \mathbf{TR}_{ma}^{pred}}{\mathbf{TR}_a^{sim}} \right| \tag{6}$$

where

e_{ma} is the absolute %error of method m for the data set a within the test set $\mathbf{T}$;

q_{T_a} is the number of variability scenarios within the test set $\mathbf{T}_a$;

$\mathbf{TR}_a^{sim}$ is the simulated TR of the scenarios within the test set $\mathbf{T}_a$; and

$\mathbf{TR}_{ma}^{pred}$ is the predicted TR using method m of the scenarios within the test set $\mathbf{T}_a$.

The results for each training experiment $\mathbf{R}_w$ were then collated to the set $\mathbf{E}$. The mean and coefficient of variation of the %errors of each method m and data set a within the test set $\mathbf{T}$, i.e., μ_e and c_e, were calculated for the set $\mathbf{E}$ to determine which method outperforms the others for the particular training set $\mathbf{R}_w$.

The error percentages were then rounded to the nearest hundredth and $\mu Score$, $cScore$ of each method m within the set $\mathbf{E}$ of the training experiment $\mathbf{R}_w$ were determined and compared to obtain the method(s) that has the smallest errors according to the scoring criteria shown in Table 2.

Subsequently, the same steps were repeated with another data set $\mathbf{D}_x$ and supporting predictor p_j used for training.

After all the possible data sets were completed, the elected set(s) $\mathbf{S}_j$ including the elected data set(s) used for training $\mathbf{S}_n$ were compiled into the optimal training set $\mathbf{R}_w$ and the covariates and regression coefficients of the best performing method within this set were extracted.

Table 2. Scoring criteria for $\mu Score$ and $cScore$.

Rounded μ_e, c_e to Hundredth	$\mu Score, cScore$
>=100%	**0**
20-99%	**1**
10-19%	**2**
9%	**3**
8%	**4**
7%	**5**
6%	**6**
5%	**7**
4%	**9**
3%	**11**
2%	**13**
1%	**15**
0%	**20**

The multiple regression model of the throughput rate of asynchronous flow line TR^{async} for the best performing polynomial regression model is expressed as:

$$
\begin{aligned}
TR^{async} = {} & \beta_1 \mu_{\max}^{-1} + \beta_2 e^{c_{av}} + \beta_3 N^{-1} + \beta_4 e^{\mu^{-1}} + \beta_5 e^{c_{av}} \mu_{\max}^{-1} + \beta_6 e^{c_{av}} N^{-1} + \\
& \beta_7 e^{c_{av}} e^{\mu^{-1}} + \beta_8 \mu_{\max}^{-1} N^{-1} + \beta_9 \mu_{\max}^{-1} e^{\mu^{-1}} + \beta_{10} N^{-1} e^{\mu^{-1}} + \beta_{11} \mu_{\max}^{-2} + \beta_{12} e^{\mu^{-2}} + \\
& \beta_{13} e^{c_{av}} \mu_{\max}^{-1} N^{-1} + \beta_{14} e^{c_{av}} N^{-1} e^{\mu^{-1}} + \beta_{15} e^{c_{av}} \mu_{\max}^{-2} + \varepsilon
\end{aligned}
\tag{7}
$$

where the model coefficients are shown in Table 3.

4. Analysis and Testing

4.1 Data Pre-processing

Eight synthetic discrete data sets were sampled. Representative data sets were defined for training and testing of the intra- and inter-variability of processing times P_i and length N within asynchronous non-exponential flow lines. The first two data sets ($I/II - A - 1$ and $I/II - A - 2$) were chosen to fully represent the processing time variability up to a scale of 10, $1 \le \mu_i \le 10$, for a relatively small flow line, $N < 5$. Data set $I/II - A - 1$ is for flow lines with lengths of one and two processes while three and four processes are covered in the data set $I/II - A - 2$.

For flow lines with one to four processes, full factorial DOE was used to generate all scenarios in the data set, where mean processing times varies between 1-10 time units.

Table 3. Regression model coefficients.

Coefficient	**Value**
β_1	1.241300
β_2	-0.074113
β_3	0.165250
β_4	0.047208
β_5	-0.385610
β_6	-0.151280
β_7	0.074081
β_8	-0.088294
β_9	0.114680
β_{10}	-0.165220
β_{11}	0.081924
β_{12}	-0.051851
β_{13}	0.120080
β_{14}	0.151210
β_{15}	-0.141977
ε	0.004667

For longer flow lines, μ_i was selected randomly and equiprobably for the second six data sets as follows:

$I/II-A-3: \mu_i \in \{1,2,...,10\}, N \in \{1,2,3,4,5\}$,

$I/II-A-4: \mu_i \in \{1,2,...,60\}, N \in \{1,2,3,4,5\}$,

$I/II-A-5: \mu_i \in \{1,2,...,100\}, N \in \{1,2,3,4,5\}$,

$I/II-A-6: \mu_i \in \{1,2,...,60\}, N \in \{1,2,...,30\}$,

$I/II-A-7: \mu_i \in \{1,2,...,100\}, N \in \{1,2,...,30\}$, and

$I/II-A-8: \mu_i \in \{1,2,...,500\}, N \in \{1,2,...,30\}$,

and $c_{av} \in \{0,0.01,0.025,0.05,0.075,0.1,0.25,0.5,0.75,1\}$ for all data sets, i.e., $I/II-A-1$ to $I/II-A-8$. Total number of experiments for all data sets is 114,093 experiments.

4.2 Feature Selection

Statistical analysis on the relationship between the parameters in Section 3.2 including their nonlinear terms and the throughput rate was carried out. Based on the results (Table 4), new nonlinear relationships between the following set of flow line-based variability parameters and TR were confirmed to a high certainty as follows:

i. The inverse of maximum mean processing time $\mu_{\max}{}^{-1}$, the coefficient of variation c_{av} and N;
ii. A nonlinear term related to the coefficient of variation, namely $e^{c_{av}}$; and
iii. One term for the length: N^{-1}.

Results also showed that although the process with maximum mean processing time has a significant effect on the throughput rate, the location of such process is irrelevant. Furthermore, parameters μ, μ^{-1}, $\log\mu$, $\log {}^{1}\!/_{\mu}$ and $e^{\mu^{-1}}$ have inconsistent relationship with the throughput rate but an acceptable statistical importance suggesting that a relationship might exist. The relationship between these terms and TR is more likely to be high for the sub-sets where there is a proximity between the average and maximum mean processing times, e.g., $\mathbf{TR}_{ss,5}$. Each parameter was given a score based on its relationship to TR. Table 5 shows the correlation coefficient for each sub-set. As shown, the correlation is strong, i.e., higher than 0.8, for two sub-sets out of 8, hence, the score given to this parameter was 2/8. The same criterion was applied to the significance of the parameter but the score was doubled. The total weighted score was then calculated and the pass score was set low, i.e., 25% or 6/24 (Table 6).

Best regression technique was applied to verify the findings. The results as shown in Table 7 suggest that main predictor terms are needed in order to define the throughput rate; exclusion of any of them has a significant effect on Mallows's Cp. Ideally, Mallows's Cp has to equals the number of predictors plus one (for the constant); this condition was met when all predictors are included. Linear regression model with the first set of parameters was accurate to a standard error of 0.0082 and R^2 of 96.7%. Adding terms of the free predictor μ has improved the accuracy and significance of the model to R^2 of 97% and 0.009 standard error. However, the term $\log {}^{1}\!/_{\mu}$, which did not show an effect on the model, can be excluded from the model building stage.

However, the best-sub-set model does not explain the relationship completely since the degrees of freedom in modelling using normal Best Regression Technique are limited (i.e., linear) which was investigated during the model building stage.

Hence, the parameters can be categorised into two categories. The main predictors with clear impact on the throughput rate are ${\mu_{\max}}^{-1}$, c_{av}, $e^{c_{av}}$, N and. The second category includes parameters with less statistically proven relationship with throughput rate. These terms, i.e., μ, μ^{-1}, $\log\mu$, and $e^{\mu^{-1}}$, were included as free predictors to improve the model accuracy.

4.3 Model Building

4.3.1 Training Set Size

The training set size can vary from a single, i.e., $n = 1$, to the total number of data sets $n = x = 8$. Decision on the number of data sets was based on the improvements achieved from each iterative step. When no further improvement could be achieved, the number of the best performing training set was maintained.

Among all data sets $\mathbf{D}_x, x \in \{1,2,...,8\}$ looped within the training data set $\mathbf{S}_n$ for the data mining models, three indexed data sets $n \in \{1,2,3\}$ with the data sets $\mathbf{D}_x$, $x = 8$, $x = 3$ and $x = 4$ showed the best improvements in the model accuracy for this DOF iterative steps. No improvements were evident with quadruple data sets. It is worth noting that addition of the wrong data set to the training set can reduce the performance.

In terms of methods and models performance, the statements in Section 4.3.1 still hold true for this iterative step. The robust fitting and regularisation algorithms of regression models were still generating high errors. Bounded pure quadratic regression without the multiplication of terms remains the worst among all stepwise regression along with polynomial regression with unbounded steps, i.e., backward iteration.

In addition, the following formulas were used for comparison purposes:

i. Deterministic throughput rate TR^d with the condition: $c_i = 0, i = 1,2,....,N$:

$$TR^d = 1/\mu_{\max} \qquad (8);$$

ii. Li and Meerkov's formula (Equation 1).

Table 4. (i) Correlation and (ii) ANOVA analysis of relationship between parameters and throughput rate.

(a) Maximum Mean Processing Time

(i)

Terms	TR_1	TR_2
$\mu_{\max}$	**-0.90**	**-0.90**
$\mu_{\max}^{-1}$	**1.00**	**1.00**
$\log \mu_{\max}$	**-0.97**	**-0.97**
$\log 1/\mu_{\max}$	**0.97**	**0.97**
$e^{\mu_{\max}}$	-0.49	-0.49
$e^{\mu_{\max}^{-1}}$	**1.00**	**1.00**

(ii)

Source	*DF*	*Adj SS*	*Adj MS*	*f-value*	*p-value*
Regression	1	0.1364	0.1364	21533270	0
$\mu_{\max}^{-1}$	**1**	**0.1364**	**0.1364**	**21533270**	**0**
Error	7	0	0		
Total	8	0.1364			

(b) Location Ratio of the Process with Maximum Mean Processing Time

(i)

Term	TR_1
l	-0.06
l^{-1}	-0.12
$\log l$	0.00
$\log 1/l$	-0.00
e^{l}	-0.07
$e^{l^{-1}}$	-0.21

(c) Coefficient of Variation

(i)

Term	TR_1	TR_2	TR_3
c_{av}	**-0.90**	**-0.97**	**-0.99**
c_{av}^{-1}	0.33	0.41	0.45
$\log c_{av}$	-0.66	-0.78	-0.83
$\log 1/c_{av}$	0.66	0.78	0.83
$e^{c_{av}}$	**-0.95**	**-0.99**	**-0.99**
$e^{c_{av}^{-1}}$	0.17	0.22	0.25

(ii)

Source	*DF*	*Adj SS*	*Adj MS*	*f-value*	*p-value*
Regression	2	0.000001	0	1256.59	0
c_{av}	**1**	**0**	**0**	**253.38**	**0**
$e^{c_{av}}$	**1**	**0**	**0**	**468.27**	**0**
Error	5	0	0		
Total	7	0.000001			

(d) Length

(i)

Terms	TR_1	TR_2	TR_3	TR_4
N	**-0.87**	**-0.88**	**-0.88**	**-0.89**
N^{-1}	**1.00**	**1.00**	**1.00**	**1.00**
$\log N$	**-0.97**	**-0.97**	**-0.97**	**-0.98**
$\log 1/N$	**0.97**	**0.97**	**0.97**	**0.98**
e^{N}	-0.30	-0.30	-0.32	-0.33
$e^{N^{-1}}$	**1.00**	**0.99**	**0.99**	**0.99**

(ii)

Source	*DF*	*Adj SS*	*Adj MS*	*f-value*	*p-value*
Regression	3	0.0034	0.0011	4925.5	0
N	**1**	**0.000002**	**0.000002**	**8.24**	**0.017**
N^{-1}	**1**	**0.00004**	**0.00004**	**149.8**	**0**
$\log 1/N$	1	0.000001	0.000001	3.59	0.087
Error	10	0.000002	0		
Total	13	0.003443			

Table 5. Correlation coefficient between the direct term of average mean processing time and the throughput rate of each sub-set.

	$\mathbf{TR}_{ss,1}$	$\mathbf{TR}_{ss,2}$	$\mathbf{TR}_{ss,3}$	$\mathbf{TR}_{ss,4}$	$\mathbf{TR}_{ss,5}$	$\mathbf{TR}_{ss,6}$	$\mathbf{TR}_{ss,7}$	$\mathbf{TR}_{ss,8}$
μ	**-0.99**	-0.75	-0.76	**-0.93**	-0.70	-0.67	-0.67	-0.67

Table 6. Relationship score and inclusion or exclusion decision for the average of mean processing times terms.

Predictor Terms	*Score Strength (out of 8)*	*Score Significance (out of 16)*	*Total Score (out of 24)*	*Decision*
μ	**2**	**4**	**6**	***Include***
μ^{-1}	**5**	**2**	**8**	***Include***
$\log \mu$	**6**	**2**	**8**	***Include***
$\log \frac{1}{\mu}$	**6**	**0**	**6**	***Include***
e^{μ}	3	0	3	*Exclude*
$e^{\mu^{-1}}$	**5**	**6**	**11**	***Include***

Table 7. Best sub-set regression analysis for (a) main predictors only and (b) main and free predictors. X indicates that the predictor is included in the regression model

(a)

Vars	*R^2*	*R^2(adj)*	*Mallows Cp*	*S*	$\mu_{\max}^{-1}$	c_{av}	$e^{c_{av}}$	N	N^{-1}
1	89.2	89.2	262821	0.0149	X				
2	96.7	96.7	1646.1	0.0083	X		X		
3	96.7	96.7	748.2	0.0082	X	X	X		
4	96.7	96.7	100	0.008	X	X	X		X
5	96.7	96.7	81.1	0.008	X	X	X		X
6	**96.7**	**96.7**	**7**	**0.0082**	**X**	**X**	**X**	**X**	**X**

(b)

Vars	*R^2*	*R^2(adj)*	*Mallows Cp*	*S*	μ	μ^{-1}	$\log \mu$	$e^{\mu^{-1}}$
1	97.0	97.0	874	0.009		X		
2	97.0	97.0	695	0.009		X		X
3	97.0	97.0	102	0.009		X	X	X
4	**97.0**	**97.0**	**11**	**0.009**	**X**	**X**	**X**	**X**

The deterministic throughput rate formula performed poorly for all test sets with $\mu Score$ and $cScore$ of 2 and 1, respectively, which suggests that variability was well introduced within the data sets. Polynomial regression model with bounded steps for this DOF increased $cScore$ to 2 which is higher than with Li and Meerkov's formula. However, the $\mu Score$ of Li and Meerkov's formula still surpasses the polynomial model. Hence, the triple training data sets were used in the next DOF iterative steps, i.e., addition of supporting predictors.

4.3.2 Supporting Predictor Terms

In this step, the supporting predictor terms were iteratively added to the training of machine learning models. The four supporting predictors are μ , μ^{-1} , $\log \mu$, and $e^{\mu^{-1}}$. The supporting predictor $e^{\mu^{-1}}$ improved the prediction accuracy such that $\mu Score$ with the bounded steps polynomial regression model reached the same as the non-standalone Li and Meerkov's formula while maintaining the $cScore$ at its higher value of 2. The best performing standalone regression model gave μ_e and c_e of 2% and 0.19 against 2% and 0.45 for Li and Meerkov's formula.

Table 8 shows the score of the average and coefficient of variation of %errors of each model, i.e., $\mu Score_m$ and $cScore_m$, for the iterative steps of the data sets used for training $\mathbf{S}_n$. Addition of two supporting predictor terms to the training data sets failed to improve the performance.

As shown in Table 8, in terms of regression models performance, robust fitting performance improved with the addition of a single supporting predictor term to become comparable to that of the stepwise regression, excluding polynomial bounded regression, while regularisation of squared errors remained a poor performer with no difference between the three algorithms with different penalties.

Figure 5 shows the predicted throughput rate of the optimal model against the simulated throughput rate for data set $I/II-A-6$ while comparing it with Li and Meerkov's formula and the best performing classification machine learning model.

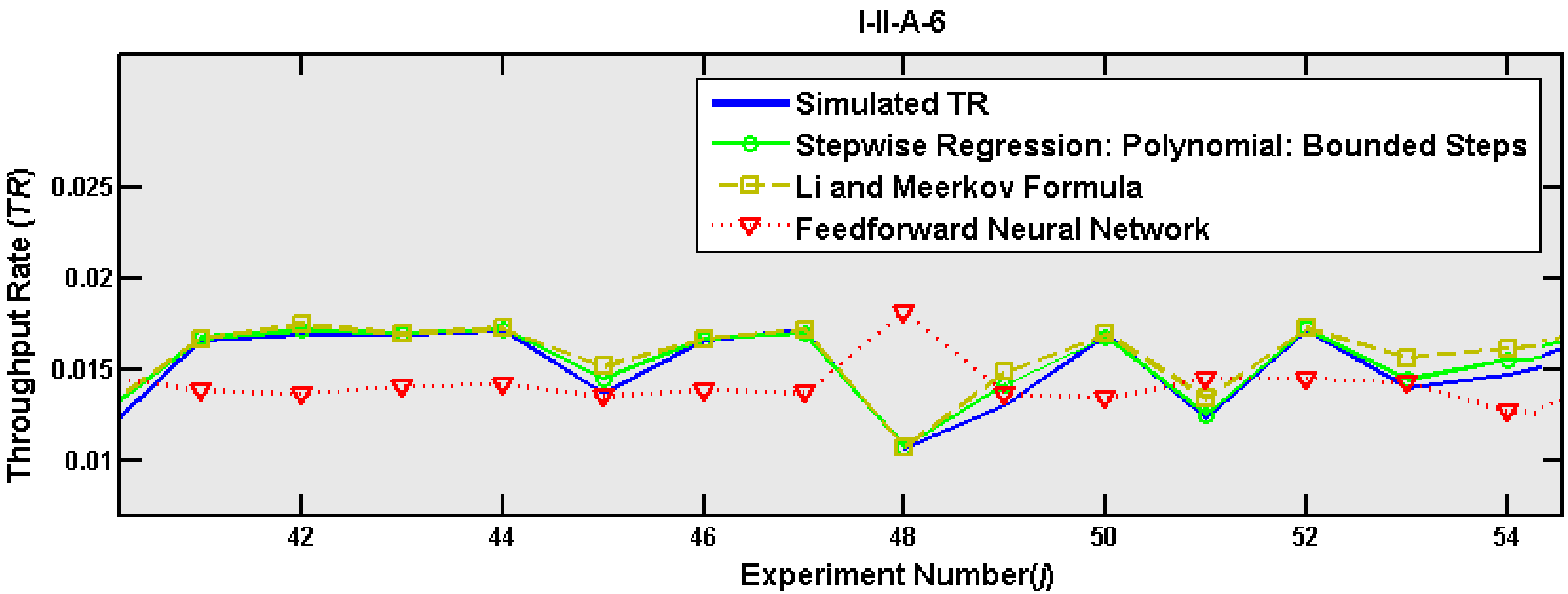

Figure 5. Throughput rate of asynchronous flow lines. Comparison of results obtained with the proposed empirical formula (Stepwise Regression), Li and Meerkov's formula [6], the best classification learning model (Feedforward Neural Network) and simulation.

Table 8. Performance of data mining models for asynchronous flow lines.

					Train Set **at n=3,i=5**		Train Set **at n=3,i=6**	
					$\log \mu$		$e^{\mu^{-1}}$	
	Method				**Score**		**Score**	
No.	**Class I**	**Class II**	**Class III**	**Class IV**	***μScore$_m$***	***cScore$_m$***	***μScore$_m$***	***cScore$_m$***
1	Li and Meerkov [6]				13	1	13	1
2	Equation 8				2	1	2	1
3	**Multiple Linear Regression**	Robust [35]	Tukey's Bisquare		4	1	4	1
4			Andrews		4	1	4	1
5			Cauchy M-estimators		4	1	4	1
6			Fair		2	1	2	1
7			Huber		2	1	3	1
8			Logistic Regression		2	1	3	1
9			Hinch and Talwar		4	1	4	1
10			Holland and Welsch		4	1	4	1
11		Regularisation [36]	Lasso		2	0	2	0
12			Ridge Regression		2	0	2	0
13			Elastic Nets		2	0	2	0
14		**Stepwise**	Linear with Interaction Effects	Bounded	2	1	6	1
15				Unbounded	2	0	7	1
16			Pure quadratic	Bounded	1	0	2	0
17				Unbounded	1	0	6	1
18			Quadratic	Bounded	5	1	6	1
19				Unbounded	2	0	2	0
20			**Polynomial**	**Bounded**	**9**	**1**	**13**	**2**

5. Case Study

5.1 Overview

The developed regression method was further used to predict the daily throughput rate of a concrete central reserve barrier (CRB) construction project under variable conditions. The case study reported here was done collaboratively with Costain Group plc, a British engineering company, and was focused on the CRB construction on UK's motorway M1 (Junction 28 to 31). The project, '*M1 Smart Motorway – Junction 28 to 31*', started in 2014 and was completed in 2016.

The daily operations start with the supplier delivering batches of concrete to the project site from two concrete plants at different locations. The trucks (6 to 8m^3) drive to the site through the motorway. At the site entrance, trucks can face delays due to site works. Once the trucks reach the project site, the concrete slump test is carried out and, based on the results, one of the following occurs:

1. Water is added to the load;
2. The truck is placed in a queue while waiting for the load to dry; or
3. The truck proceeds to the discharge process if the extruder is free.

Once the load is discharged, the operation is considered complete. A saw-cut process of the barrier takes place after the discharge process. However, this happens independently so it does not affect the completed barrier length.

Figure 6 shows a simulation model of the CRB project. The main elements of the model are:

a. Work Entry Point: where trucks enter the system before any processing operation is initiated;
b. Batch and Load Queue (Q1): the queue of concrete trucks waiting to be batched;
c. Batch and Load Process: the first process at the contractor concrete plant sites, where trucks are batched and loaded with concrete;
d. Drive to Site Process: the second process, where trucks are delivering concrete to the construction site and can be delayed due to traffic congestion;
e. Site Access Queue (Q2): delay to the concrete deliveries at the site access.
f. Site Queue (Q3): trucks waiting to be load-tested at the site.
g. Slump Test Process: the slump test is applied to the concrete load to determine its suitability;
h. Add Water Process: in case of dry load;
i. High Slump Load Queue (HSLQ): where trucks wait for the high slump load to dry;
j. Discharge and Extrude Process: the only value-added process, where the load is being discharged at site;
k. Saw Cut Process: an additional process that occurs after the barrier is extruded; and
l. Work Exit Point: where items leave the system.

Multiple variables within this project cause disruption to the concrete deliveries and consequently to the completed barrier length, i.e., throughput rate. The objective is to accurately estimate the anticipated throughput rate using the developed regression model, taking into consideration the variability factors that play a part in the construction operations and the different constraints and operational conditions during the working day, e.g., traffic congestion.

Standalone closed-form formula of the throughput rate

5.2 Developed Model Validation

Concrete pour records for the CRB construction project were collected for two months of operations and processed to obtain the variability within the construction project (Table 9):

i. Arrival rate, i.e., schedule of deliveries;
ii. Concrete batch time;
iii. Truck delivery time;
iv. Pre-test site delays time;
v. Load conditioning time; and
vi. Truck discharge time.

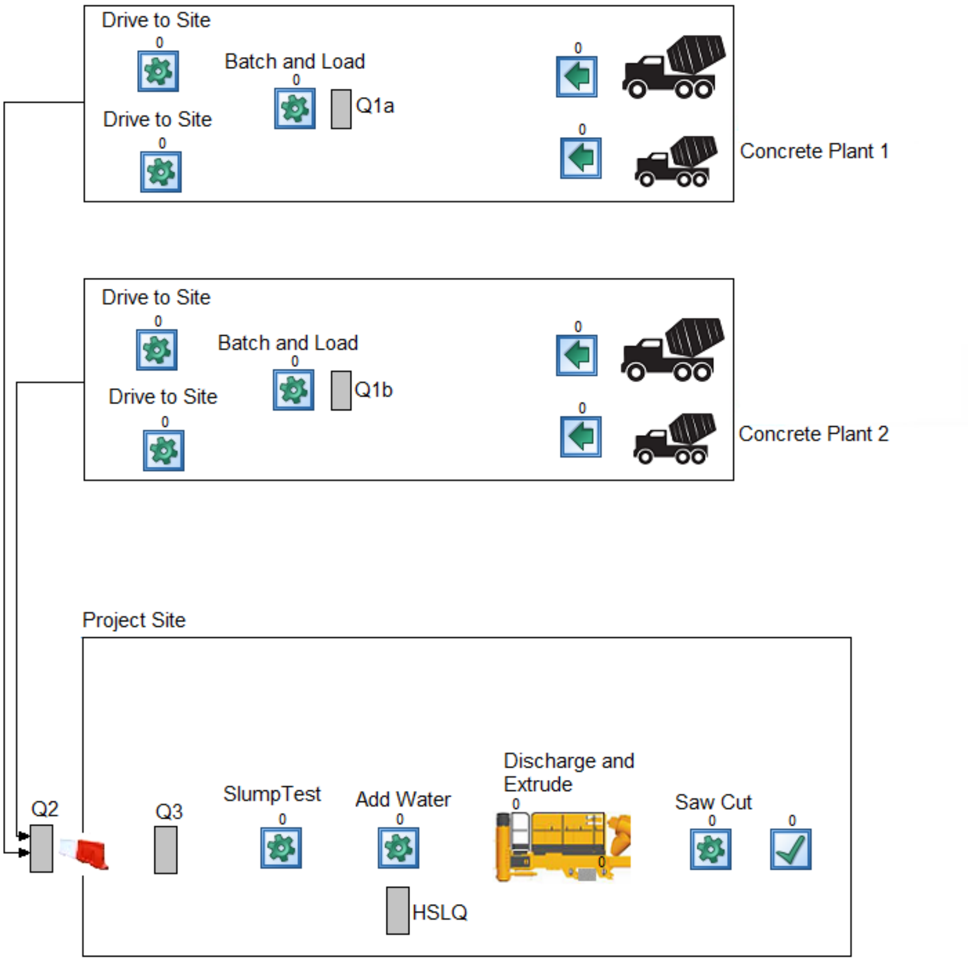

Figure 6. Simulation model of the case study.

As shown in Table 9, the truck load size affects the batch and discharge processing time. Concrete plant location and traffic congestions decide the delivery time. In terms of '*Load Conditioning Time*', three load conditions can be expected at the project site:

i. Suitable load: ready to be discharged when the extruder machine is available;

ii. High slump: the truck needs to wait for the concrete load with high water content to dry; and

iii. Low slump: water is added to the load to adjust concrete properties.

Variability factors ii to vi are generally uncontrollable since Costain has minimal control on the choice of concrete plant and truck sizes. However, the arrival rate (variability factor i) is highly controllable. Combinations of all possible variability scenarios were generated based on historical data and simulations were applied using Simul8 software with a confidence interval of 95% to determine the anticipated throughput rate for each scenario. Figure 7 shows a comparison between the simulated and calculated throughput rate using the developed model.

Table 10. Variability of the case study.

Arrival Rate		
Number of Trucks	$1/\mu_{ar}$ *(1/min)*	c_{ar}
1	5-80	0-1

Concrete Batch Time		
Load Size (m^3)	μ_b *(min)*	c_b
1	2	0.25

Truck Delivery Time				
Concrete Plant	P1		P2	
Time-of the-day	μ_{del} *(min)*	c_{del}	μ_{del} *(min)*	c_{del}
07:00	34.10	0.26	51.80	0.52
08:00	38.03	0.37	50.27	0.24
09:00-15:00	30.55	0.28	48.96	0.24
16:00	29.44	0.26	39.00	0.35

Pre-test site delays time		
Number of Trucks	μ_s *(min)*	c_s
1	3.24	0.22

Load Conditioning Time		
Load Condition	μ_c *(min)*	c_c
Low Slump	11.14	0.49
High Slump	36.47	0.31

Truck Discharge Time		
Load Size (m^3)	μ_{dis} *(min)*	c_{dis}
1	4.39	0.35

The results show that the calculated *TR* follows closely the simulated *TR* with a margin of error of ±5%. The increase in error and presence of oscillations can be linked to the introduction of continuous data in the validation case study which were not present in the test sets $(I/II-A-1$ to $I/II-A-8)$. Weak correlation between the residuals and the model variables (Table 10) suggests that the model performance remains valid to other case studies with different continuous data.

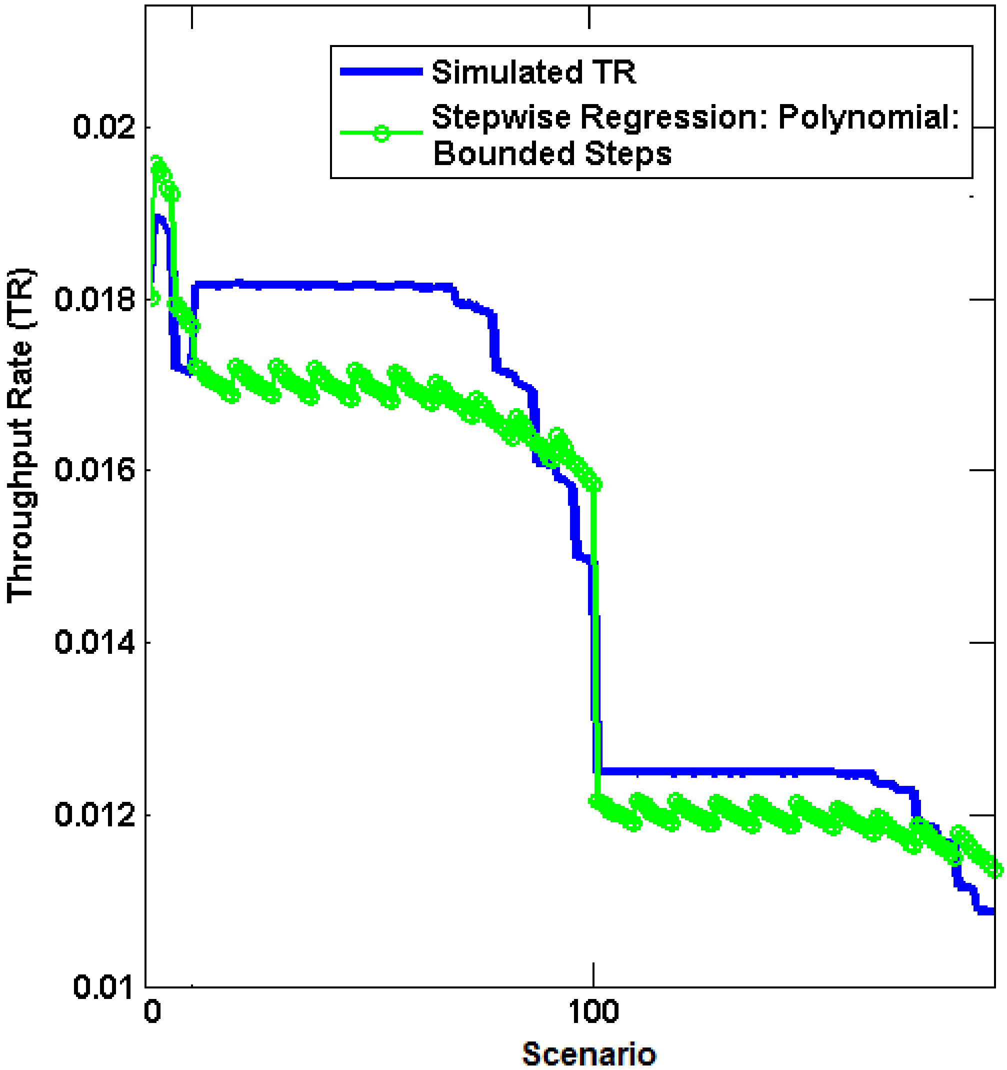

Figure 7. Throughput rate of case study. Comparison of results obtained with the proposed model (stepwise regression) and simulation.

Table 10. Correlation coefficient between residuals and the proposed model's variables.

	${\mu_{\max}}^{-1}$	$e^{c_{av}}$	$e^{\mu^{-1}}$
Residuals	-0.25	-0.15	-0.31

6. Conclusion

This paper proposed a closed-form empirical evaluative model to easily and quickly estimate the effect of stochastic variability in process and production planning on the system-level performance of flexible human-dependent serial flow lines.

Through this investigation, the following main contributions were achieved:

1. Generic representation of arbitrary length non-exponential serial flow lines using nonlinear terms. New nonlinear relationships between the normal distribution-based variability parameters and TR were identified with *p*-values less than 0.01 and correlation coefficients higher than 0.8; and
2. Standalone closed-form empirical formula that estimates the throughput rate of asynchronous flow lines with normally distributed process variability to a higher accuracy and independency than existing formulas. The best performing standalone regression model with the optimum training set (Equation 7) gave the same average prediction error as the non-standalone formula of Li and Meerkov's [6] but with an improved coefficient of variation of prediction error (0.19 against 0.45).

The paper mainly focused on normal distributions. However, Li and Meerkov [6, p. 549] showed that the throughput rate is not very sensitive to the distribution type and is almost a linear function of the average coefficient of variation if the coefficient of variation is smaller than one. Future work will try to develop an autonomous control method based on the developed empirical formula for asynchronous flexible flow lines and validate it in a real-world environment using continuous data.

Acknowledgements

This work was supported by InnovateUK [Grant Number: 18834-132285].

References

[1] He XF, Wu S, Li QL. Production variability of production lines. Int J Prod Econ 2007;107:78–87. doi:10.1016/j.ijpe.2006.05.014.

[2] Ambani SM. Analytical estimation of throughput distribution for serial manufacturing systems with multi-state machines and its application. University of Michigan, 2011.

[3] Wang J, Zhong X, Li J. Modelling and analysis of care delivery services within patient rooms: A system-theoretic approach. IEEE Trans Autom Sci Eng 2014;11:379–93.

[4] Oecd. Understanding the Brain: The Birth of a Learning Science. 2007. doi:10.1787/9789264029132-en.

[5] Carrascosa M. Variance of the output in a deterministic two-machine line. Massachusetts Institute of Technology, 1995.

[6] Li J, Meerkov SM. Production systems engineering. New York: Springer; 2009.

[7] Spinellis D, Vidalis MJ, O'Kelly MEJ, Papadopoulos CT. Analysis and Design of Discrete Part Production Lines. Computer (Long Beach Calif) 2009;31:1–23. doi:10.1007/978-0-387-89494-2.

[8] Li J, Meerkov SM, Zhang L. Production systems engineering: Review and recent developments. Handb. Stoch. Model. Anal. Manuf. Syst. Oper., New York: Springer; 2013, p. 167–210.

[9] Tan B, Gershwin SB, Liberopoulos G, Meerkov SM, Papadopoulos CT. Advances in stochastic models of manufacturing and service operations. Ann Oper Res 2015;231. doi:10.1007/s10479-015-1868-7.

[10] Meerkov SM, Yan C. Production lead time in serial lines: evaluation, analysis, and control. IEEE Trans Autom Sci Eng 2014;99:1–13.

[11] El-Haik B, Al-Aomar R. Simulation-Based Lean Six-Sigma and Design for Six-Sigma. 2006. doi:10.1002/0470047720.

[12] Hopp WJ, Spearman ML. Factory Physics: foundation of manufacturing management. 2008.

[13] Baykasoğlu A, Gocken M. A simulation based approach to analyse the effects of job release on the performance of a multi-stage job-shop with processing flexibility. Int J Prod Res 2011;49:585–610. doi:10.1080/00207540903479778.

[14] Heilala J. Use of simulation in manufacturing and logistics systems planning. Proc. VTT Manuf. Technol., Helsinki: 1999, p. 1–24.

[15] Scholz-Reiter B, Freitag M. Autonomous Processes in Assembly Systems. CIRP Ann - Manuf Technol 2007;56:712–29. doi:10.1016/j.cirp.2007.10.002.

[16] Sethi AK, Sethi SP. Flexibility in manufacturing: A survey. Int J Flex Manuf Syst 1990;2:289–328. doi:10.1007/BF00186471.

[17] Wiendahl HP, ElMaraghy HA, Nyhuis P, Zah MF, Wiendahl HH, Duffie N, et al. Changeable Manufacturing - Classification, Design and Operation. CIRP Ann - Manuf Technol 2007;56:783–809. doi:10.1016/j.cirp.2007.10.003.

[18] Windt K, Jeken O. Allocation Flexibility - A New Flexibility Type as an Enabler for Autonomous Control in Production Logistics. 42nd CIRP Conf Manuf Syst 2009:8.

[19] Li J, E. Blumenfeld D, Huang N, M. Alden J. Throughput analysis of production systems: recent advances and future topics. Int J Prod Res 2009;47:3823–51. doi:10.1080/00207540701829752.

[20] Leu YY, Matheson LA, Rees LP. Sequencing mixed-model assembly lines with genetic algorithms. Comput Ind Eng 1996;30:1027–36. doi:10.1016/0360-8352(96)00050-2.

[21] Buhne S, Lauenroth K, Pohl K. Modelling requirements variability across product lines. Requir Eng 2005 Proceedings 13th IEEE Int Conf 2005:41–50. doi:10.1109/RE.2005.45.

[22] Trzyna D, Kuyumcu A, Lödding H. Throughput time characteristics of rush orders and their impact on standard orders. Procedia CIRP, vol. 3, 2012, p. 311–6. doi:10.1016/j.procir.2012.07.054.

[23] Martin WE, Bridgmon KD. Quantitative and statistical research methods From hypothesis to results. 2012. doi:10.1017/CBO9781107415324.004.

[24] Hopp WJ. Supply Chain Science. Long Grove: Waveland Press; 2008.

[25] Inman RR. Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. Prod Oper Manag 1999;8:409–32. doi:10.1111/j.1937-5956.1999.tb00316.x.

[26] Blumenfeld, D. E. & Li J. An analytical formula for throughput of a production line with identical stations and random failures. Math Probl Eng 2005;3:293–308.

[27] Hunt GC. Sequential Arrays of Waiting Lines. Oper Res 1956;4:674–83. doi:10.1287/opre.4.6.674.

[28] Miltenburg GJ. Variance of the number of units produced on a transfer line with buffer inventories during a period of length t. Nav Res Logist 1987;34:811–22.

[29] Gershwin SB. Manufacturing Systems Engineering. New Jersey: Prentice-Hall; 1994.

[30] Kang N, Zheng L, Li J. Analysis of multi-product manufacturing systems with arbitrary processing times. Int J Prod Res 2015;53:983–1001.

[31] Papadopoulos HT. An analytic formula for the mean throughput of K-station production lines with no intermediate buffers. Eur J Oper Res 1996;91:481–94. doi:10.1016/0377-2217(95)00113-1.

[32] Jacobs JH, Etman LFP, Van Campen EJJ, Rooda JE. Characterization of operational time variability using effective process times. IEEE Trans Semicond Manuf 2003;16:511–20. doi:10.1109/TSM.2003.815215.

[33] Li J, Meerkov SM. Due-Time Performance of Production Systems with Markovian Machines. Anal. Model. Manuf. Syst., Dordrecht: Kluwer Academic; 2003, p. 221–53.

[34] Mendenhall W, Beaver R, Beaver B. Introduction to Probability and Statistics. Boston: Cengage Learning; 2012.

[35] MathWorks. Robust Regression. MathWorks Doc 2016. http://uk.mathworks.com/help/stats/robustfit.html (accessed May 31, 2016).

[36] MathWorks. Lasso and Elastic Net. MathWorks Doc 2016. http://uk.mathworks.com/help/stats/lasso-and-elastic-net.html (accessed May 31, 2016).

Appendix A

N	Number of processes
$\boldsymbol{\beta}$	Regression coefficients
c	Coefficient of variation
c_{av}	Average coefficient of variation
$cScore$	Score of the errors in c
$\mathbf{D}$	Data set
e	Mean absolute percentage error
$\mathbf{E}$	Mean absolute percentage error matrix
ε	Regression error term
i	Location of process
j	Number of variability scenarios
m	Machine learning method counter
l	Location ratio of the process with maximum mean processing time
μ	Mean processing time
$\mu_{\max}$	Maximum mean processing time
$\mu_{\min}$	Minimum mean processing time
$\mu Score$	Score of the errors in μ
n	Counter of data sets in training set
P	Process
$\mathbf{R}_w$	Training experiment
S	Train set
ss	sub-set
$\mathbf{T}$	Test set
TR	Throughput rate
TR^d	Throughput rate of deterministic flow line
w	Training set number
χ	Data set counter
y	Predictor counter

Adam Aboutaleb is a Lecturer in Engineering Management and Innovation at Shaoyang University, Hunan, China. He completed his MSc degree in Mechatronics and PhD degree in Manufacturing Science at De Montfort University (DMU), Leicester, UK, in 2012 and 2016, respectively. He was a Research Fellow at the Advanced Manufacturing Processes and Mechatronics Centre within DMU between 2012 and 2015 managing and implementing several Innovate UK projects in manufacturing, material science, construction and mechanical engineering. His research targets evaluative modelling and autonomous control of production lines. He also has 4-year industrial experience in equipment maintenance and project engineering for multi-million pound projects in container terminals.

Dr Parminder Singh Kang holds Senior Visiting Research Fellow position at De Montfort University (DMU), Leicester, UK. He received B-Tech degree (Computer Science and Engineering) in 2006 form Punjab Technical University, India. He received M.Sc. degree (IT) in 2008 and completed his PhD in Manufacturing Science in 2012 at DMU, UK. He has over 6 years of experience working on externally funded collaborative industrial R&D cross-disciplinary project in the area of organizational operations optimization. His research interests are; evolutionary algorithms, combinatorial optimization, simulation modeling, Lean/Six Sigma and application of these techniques (integrated approaches) in industrial/service process improvement and operations optimization.

Raouf Hamzaoui received the MSc degree in mathematics from the University of Montreal, Canada, in 1993, the Dr.rer.nat. degree from the University of Freiburg, Germany, in 1997, and the Habilitation degree in computer science from the University of Konstanz, Germany, in 2004. He was an Assistant Professor with the Department of Computer Science of the University of Leipzig, Germany and with the Department of Computer and Information Science of the University of Konstanz. Since 2006, he has been a Professor in Media Technology in the School of Engineering and Sustainable Development at De Montfort University, Leicester, UK.

Alistair Duffy is Professor of Electromagnetics at De Montfort University (DMU), Leicester, UK. He received his BEng (Hons) and MEng degrees in 1988 and 1989, respectively, from University College, Cardiff, Wales. He read for his PhD at Nottingham University, graduating in 1993. He also holds an MBA from the Open University, UK, graduating in 2004. He is a Fellow of the IEEE and the Institution of Engineering and Technology (IET, formerly IEE). He published over 200 papers, mostly on his research interests of validation of computational electromagnetics; physical layer components, particularly communications cabling, and electromagnetic compatibility testing and technology management.