Production & Manufacturing Research

An Open Access Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tpmr20

Generation of synthetic manufacturing datasets for machine learning using discrete-event simulation

K. C. Chan, Marsel Rabaev & Handy Pratama

To cite this article: K. C. Chan, Marsel Rabaev & Handy Pratama (2022) Generation of synthetic manufacturing datasets for machine learning using discrete-event simulation, Production & Manufacturing Research, 10:1, 337-353, DOI: 10.1080/21693277.2022.2086642

To link to this article: https://doi.org/10.1080/21693277.2022.2086642

PRODUCTION & MANUFACTURING RESEARCH
2022, VOL. 10, NO. 1, 337–353
https://doi.org/10.1080/21693277.2022.2086642

Taylor & Francis
Taylor & Francis Group

Generation of synthetic manufacturing datasets for machine learning using discrete-event simulation

K. C. Chan[a,b], Marsel Rabaev[b] and Handy Pratama[b]

[a]School of Business, University of Southern Queensland, Springfield, Australia; [b]School of Mechanical & Manufacturing Engineering, The University of New South Wales, Sydney, Australia

ABSTRACT
Recent advances in computing power have seen machine learning becoming an area of significant interest in manufacturing for scholars attempting to realise its full potential. Successful machine learning applications require a great amount of specific production data that is not easily nor publicly accessible. This study aims to develop a framework to use discrete-event-simulation (DES) to generate large datasets for training machine learning models. Three DES models were designed and executed to generate synthetic production data for different manufacturing scenarios. Inferences were made on the dependency between the time required to generate data and the complexity of the simulation model. The experimental results show that with the incremental changes in the simulation model, the time required to generate synthetic data tends to increase. The study revealed that DES is an effective tool for generating high-quality synthetic data which can be fed into machine learning models for training. The datasets generated by the simulations are made publicly available.

ARTICLE HISTORY
Received 1 August 2019
Accepted 28 May 2022

KEYWORDS
Discrete-event simulation; manufacturing process modelling; synthetic data generation; machine learning dataset

1. Introduction

In order to successfully apply machine learning, there must be sufficient datasets for machine learning model training (Ge et al., 2017; Wuest et al., 2016). For manufacturing problems, the data can either be collected from real-world production systems or virtual models of the production systems through DES. If field data (or real data) can be collected, researchers can use them to train machine learning models. Alternatively, researchers have to generate synthetic data through DES to overcome the lacking of data (Denkena et al., 2014; Lechevalier et al., 2015; Zhang et al., 2018). The synthetic data generated by DES can be used independently or combined with real data for machine learning (ML) model training.

There are many examples of ML applications in manufacturing using real and/or synthetic data. These applications have been reported in a wide range of manufacturing problems such as quality control (Weimer et al., 2016; Wuest et al., 2014), demand forecasting (Bajari et al., 2015; Lee et al., 2014), condition monitoring (Azadeh et al., 2013; Janssens et al., 2016), job shop scheduling (Shahzad & Mebarki, 2012), lead time

CONTACT K. C. Chan kc.chan@usq.edu.au School of Business, University of Southern Queensland, Springfield, QLD, Australia

prediction (Pfeiffer et al., 2016), supply chain capacity prediction (Silva et al., 2017), and bottleneck analysis (Subramaniyan et al., 2020, 2018). The use of synthetic data has also been reported in a number of production scheduling studies (Kaylani & Atieh, 2016; Kritzinger et al., 2018; Weichert et al., 2019).

This paper proposes a step-by-step framework that uses DES to generate synthetic datasets for the training and testing of ML models. The datasets generated are generic and comprehensive so they can be used for multiple manufacturing problems. In this study, three examples of manufacturing systems were simulated, and the performance of the synthetic data generation process for each test study was evaluated. The full datasets are made publicly available through the cloud-based repository Mendeley Data (https://data.mendeley.com/datasets/3rw227zxt7/2), with a view to freely disseminate for future research in solving manufacturing problems using machine learning.

2. Existing research work

This section briefly reviews the implementation of DES to generate synthetic data for manufacturing problems and describes a research gap identified in the field of machine learning data generation research.

Because DES can simulate the randomness or uncertainty of events or processes, the generated synthetic data closely resembles the real data. Koh and Saad (2003) proved this by simulating multiple levels of dependent demands, and then generating data on how delayed part delivery and delayed finished product delivery were affected due to the uncertainty that occurred in the production processes. Maas and Standridge (2005) combined real data and simulation generated data for capacity analysis, resource allocation, scheduling, and inventory control. Huang et al. (2013) used DES to analyse the rescheduling problem for a more complex mixed-line production facility. Zhuo et al. (2012) utilised DES to improve space allocation for block assembly. Gyulai et al. (2014) argued that the data generated by simulation can be used as input to the same model for different production schedules. Their results showed that DES can be utilized as a tool to create a robust production and capacity control for flexible assembly lines. Nyemba and Mbohwa (2017) used DES to achieve higher throughput for a multi-product assembly plant and concluded that DES was useful for the company's production planning and scheduling.

In these synthetic data generation studies, the amount of data generated is usually small. Although many applications use DES to generate data, the number of applications that use synthetic data for big data analysis is limited (Greasley & Edwards, 2021). For examples, Priore et al. (2006) utilised DES to generate a modest size of 1000 data points for ML training to solve a dynamic scheduling problem; Shiue (2009) generated only 2000 data points to train the ML algorithm for shop floor control.

Spanning across multiple areas of DES, big data, and AI/ML in production research, synthetic data generation plays an increasingly important role that warrants more focused research attention. In data mining, the CRISP-DM (Cross Industry Standard Process for Data Mining) project specifies a comprehensive process model for conducting data mining projects. The process model is independent of both the industry sector and the technology used (Azevedo & Santos, 2008; Schröer et al., 2021; Wirth & Hipp, 2000). Although such a process model is available, most studies reported in the literature do not follow any specific process model. Similarly, our literature review shows that there is no standard model or framework exists for the synthetic data generation process. Therefore, this study aims to propose a structured framework for synthetic data generation using DES for manufacturing problems.

3. Proposed framework

The proposed synthetic data generation framework is described step by step as follows:

(1) Define the layout and demand behaviour of the manufacturing plant.

(2) According to the defined manufacturing layout, use any DES software to construct the DES model.
  - In our test studies, the software ARENA was used.

(3) Use the constructed simulation model to generate manufacturing process data.
  - The dimension of the data is dependent upon the complexity of the manufacturing layout. For example, in the first test study (simplest manufacturing layout modelled), nine dimensions of data are written, while in the third test study (most complex manufacturing layout), 77 dimensions of data are written.
  - Examples of process data include:
    - The demand of each SKU within the defined specific time frame
    - The utilization of each facility and number of each SKU produced by each facility
    - Time (Value-added time, Waiting time, Non-Value-added time, other time).

(4) Record the time required to generate data.
  - In this study, the simulation time is available and can be retrieved from the metadata of the data files generated by ARENA.
  - This time information is necessary for planning the simulation runs and estimating the time requirements.

(5) Plan the execution of experiments to obtain the full dataset.
  - The full dataset will consists of repetitions of experiments using combinations of parameter ranges. Based on the data generation time obtained from the initial trials, plan to run experiments. Depending on the computing resources available, the experiments can be run in sequence or in parallel when multiple physical or cloud systems are available.

(6) According to the plan, run the simulation experiments and save the complete dataset in files with all data points recorded.

(7) Check, clean, combine and store the full dataset.
  - At this point, the synthetic data generation process is considered complete. The dataset is now ready for the next phase of training machine learning algorithms, or big data analysis and/or visualisation.

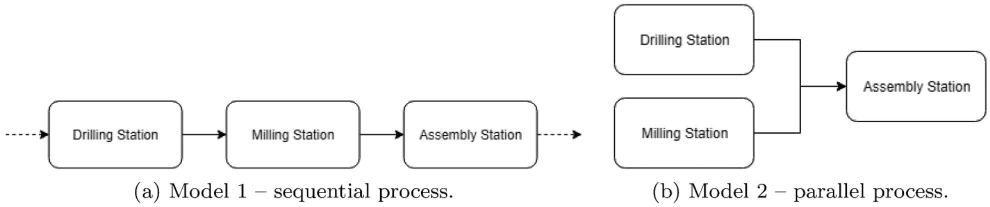
4. Test studies

When simulating a dynamic system, a flowchart is often used to deconstruct the sequence of operations, particularly to define the elements which are within and beyond the scope of a simulation (Allen, 2011; Van der Zee & Van der Vorst, 2007). The scope of this research includes three DES models – sequential process, parallel process, and flexible manufacturing, as shown in Figures 1 and 2. These models share a commonality of mixing deterministic and stochastic modelling. Some input values are manually assigned, and some are randomly generated. One of the most important randomly generated inputs is demand, which imitates the daily variation (uncertainty) of consumption levels and changes according to a uniform distribution. The distribution parameters of the first, second, and third layouts are shown in Table 1. In addition, the demand occurrence time follows an exponential distribution. Assuming that all demand values have the same probability of occurrence, a uniform distribution is selected. Finally, a replication length of 24 hours and a warm-up period of zero are set.

4.1. DES model 1 – sequential process

4.1.1. DES model 1 – description

The first model represents a simple sequential manufacturing line with three processes: drilling, milling, and assembly. The layout imitates a simple production line where blanks arrive at the first station and require drilling operations. The drilling station is capable of processing up to fifteen blanks simultaneously. The processing time follows a triangular distribution with parameters given in Table 2. If there is no available machine (resource)



(a) Model 1 – sequential process. (b) Model 2 – parallel process.

Figure 1. Layout of model 1 and 2.

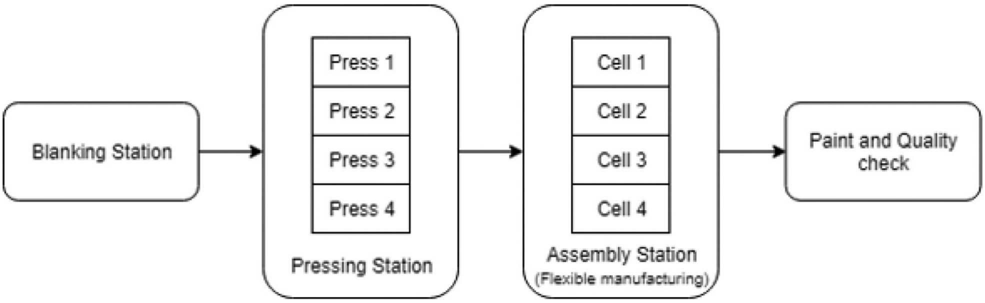


Figure 2. Layout of model 3 – flexible manufacturing cells.

Table 1. Demand parameters by layout and product type.

Layout	Product type	Minimum	Maximum
Sequential Process	Part 1	1 part per EXPO(3) min	20 parts per EXPO(2) min
Parallel Process	Part 1 & Part 2	1 part per EXPO(2) min	20 parts per EXPO(2) min
Flexible Manufacturing	SKU1, 2, 3, & 4	200 parts per 5 min	230 parts per 5 min

Table 2. Process parameters of the first layout.

	Process time (min)			
Station	Min	Mode	Max	Number of resources
Drilling	2	3	4	15
Milling	2	3	4	20
Assembly	2	3	4	10

online, the part will wait in a queue, which is an infinite size buffer for simplification. A similar approach is taken in the second and third stations. The drilled part moves to the next station where milling is performed. When finished, it moves on to the assembly process, where additional standard parts are used, which is out of the scope of the model. It should be noted that this simplified model assumes that the movement of parts between stations does not require time (nor changeover time), the service life of the machines is unlimited, and all queues are infinite. As shown in Table 2, each station follows the same distribution with the same parameters. The differences in utilisation between stations were achieved by assigning different numbers of available machines. For example, the assembly station was intentionally bottlenecked by assigning the lowest number of available machines.

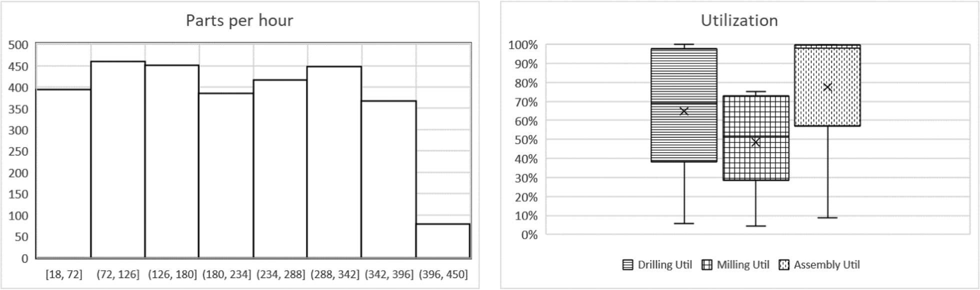
For each DES model, the data features and time requirements for generating the synthetic dataset will be described. For consistency, all simulation experiments were conducted on a Microsoft Windows PC with an Intel CPU (4 Cores 3.8 GHz) and 16 Gigabytes of Random Access Memory (RAM).

4.1.2. DES model 1 – data features

The synthetic dataset of model 1 consists of nine features of data, collected during simulation runs:

- Demand as number of parts counted at the end of replication – 1 feature (part counter).
- Average number of parts counted per hour throughout the replication – 1 feature (part counter).
- Average of total value-added time throughout the replication – 1 Feature (time).
- Waiting time per process (drilling, milling, assembly) – 3 Features (time).
- Utilisation per process (drilling, milling, assembly) – 3 Features (utilisation).

Demand is assigned on a random basis and assumed to be fulfilled by daily production (1 replication or 24 hours). The average part per hour is written as a validation tool for the model as this feature will be monitored during a simulation run. The average total value-added time shows how long a part will be processed at all the stations combined. The value-added time can be written for each station or process separately. However, since process time is set to be identical for all three stations, and the layout is a series system, the value-added time is instead written as total time, while waiting time is the average time a part waits in queue before it is further processed. Since all stations differ in resource quantities, waiting time is written separately for each station. Utilisation is the proportion of resources occupied, averaged throughout the replication, to show how busy a process is.



(a) Model 1 – histogram of parts per hour.

(b) Model 1 – box plot of utilsation.

Figure 3. Model 1 – production rate and machine utilisation.

The results of the data generation stage can be shown as distributions of the generated synthetic data. The total number of products produced per day (or in this case equal to demand) and the utilisation of each station are shown in Figure 3. The capacity of the milling station is the highest; therefore, its utilisation is the lowest, whereas the capacity of the assembly station is the lowest; therefore its utilisation is the highest (near 100%). The number of parts produced per hour tends to follow a uniform distribution, so the same is true for utilisation. These results showed that the simulation model is valid and the generated data covers most possible production scenarios, which is valuable for training machine learning models. These findings are important as they demonstrate the ability of DES to generate synthetic data for training machine learning algorithms.

4.1.3. DES model 1 – data generation time requirement

The next step in the analysis is to investigate the time required for generating data for the DES model. The simulation time can be obtained from the metadata of the generated synthetic data. In each experimental run, synthetic data is generated from 60,000 replications. The time to generate data is aggregated for every 1000 replications, so there are 60 data points in a simulation run. The time required for generating data for each model is shown in Figure 4. For model 1, in each experiment run, it was observed that the time required to generate an additional 1000 replications increased by 2.716 seconds. This phenomenon is suspected to be due to the accumulation of computing resources during the simulation process, resulting in a decrease in available memory. The time required to generate data is dependent on the process sequences, and the complexity of the manufacturing layout. This will be discussed further in the time requirement section of models two and three. A summary of the time parameters is provided in Table 3.

4.2. DES model 2 – parallel manufacturing process

4.2.1. Simulation model 2 description

The second layout contains the same three stations but is organised differently. Both drilling and milling work as an independent parallel process. The stations receive their own blanks – Part 1 and Part 2 respectively.

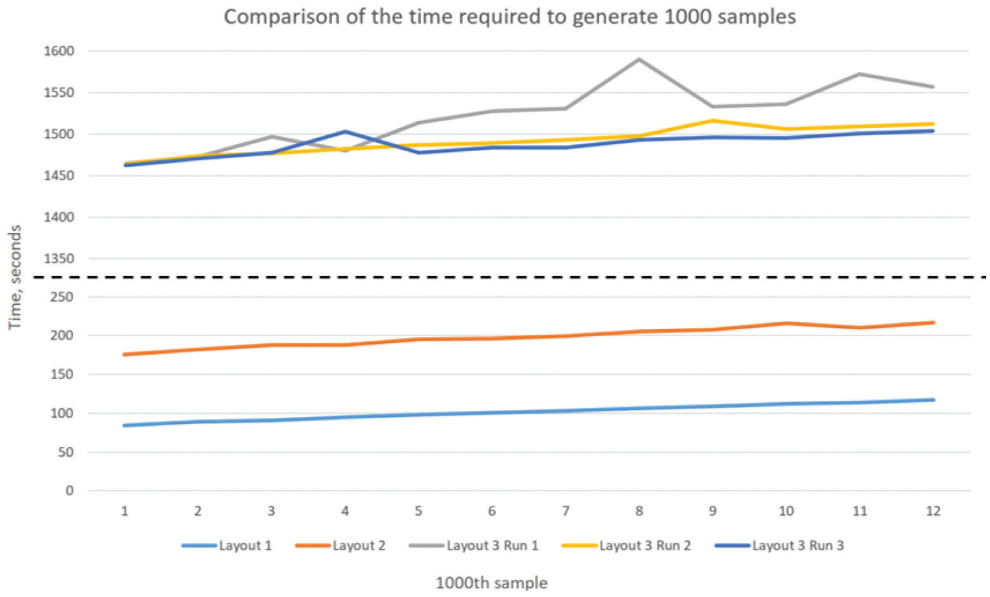


Figure 4. Time requirements to generate synthetic data for all three models.

Table 3. Model 2 – summary of time requirement parameters for data generation.

Parameter of Time Requirement	Model 1
Total numbers of data generated in the run	60000
Amount of sequences of 1000 data	60
Minimum time to generate 1000 data in the run	84 Seconds
Maximum time to generate 1000 data in the run	247 Seconds
Average time to generate 1000 data in the run	167.20 Seconds
Average time to generate 1000 data (first 12 sequences)	101.58 Seconds
Gradient – average increase of time every 1000 data in the run	2.716 Seconds
Gradient – average increase of time every 1000 data (first 12 sequences)	2.906 Seconds

Similar to the first test study, the processing time follows a triangular distribution. When there are not enough machines to process all parts, they wait in a queue before the next corresponding station. Once the parts are processed, they are sent to two corresponding infinite size buffers and held until the next station is ready and available. Both parts are picked up and moved to the assembly station by a picker. Some assumptions about the work of pickers are:

- Both drilled and milled parts must be available to be picked up, otherwise, the picker will wait.
- At least one machine must be free at the assembly station, otherwise, the picker will wait.
- No time is required to pick up and deliver parts.
- It takes 10 seconds for the picker to return to the starting point.

Table 4. Process parameters of the second layout.

	Process time (min)			
Station	Min	Mode	Max	Number of resources
Drilling	2	3	4	15
Milling	2	3	4	20
Assembly	2	3	4	10
Picker	–	0.2	–	1

The assembly process also follows a triangular distribution. However, in contrast to the previous layout, parts do not queue before the assembly station but are held in the warehouse. Hence, it is possible to track the statistics of each part, such as value-added time. Detailed information about processing times and resources is consolidated in Table 4. The expanded flowchart of the layout can be found in the Mendeley Data repository.

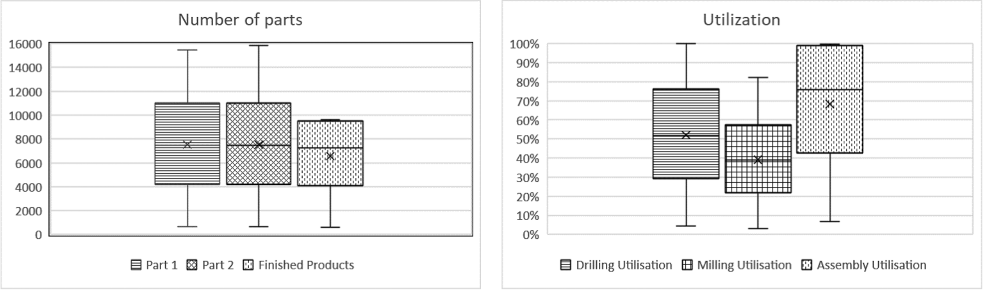
The second model layout is similar to the first model but with a different workflow. This is to allow a comparison between the models, and particularly to investigate how each performs differently.

4.2.2. DES model 2 – data features

The synthetic dataset of model 2 consists of 16 features of data, collected during simulation runs:

- Total part produced in each station – 3 features (part counter).
- Average inventory level or total part stored in drilling and milling warehouses – 2 features (part counter).
- Average value-added time for each station – 3 features (time).
- Average waiting time for each station – 3 features (time).
- Average time part spent in drilling and milling warehouses only – 2 features (time).
- Average utilization for each station – 3 features (utilisation).

Most of the data features are defined in the same way as in the first model. The new features are the part counters, which are used to record the average inventory levels of the drilling and milling warehouses; and the time features, which are used to record the average time that a part stored in the drilling and milling warehouses. The total number of parts produced per day and the utilisation of each station are shown in Figure 5. Since the demand parameters are equal, the two distributions of parts produced per day are similar. However, the distribution of finished products produced per day is expected to be different from the distributions of parts. The slight difference is due to the fact that some unfinished parts are stored in the warehouse at the end of each replication; however the measures of central tendency are similar. The utilisation results are also in line with expectations. Similar to model 1, the milling process exhibited the lowest utilisation due to its highest capacity. The simulation itself was valid, and no outliers, missing values, or any other inconsistencies were observed.



(a) Model 2 – box plot of number of parts produced per day.

(b) Model 2 – box plot of utilsation.

Figure 5. Model 2 – parts produced per day and machine utilisation.

Table 5. Model 2 – summary of time requirement parameters for data generation.

Parameter of Time Requirement	Model 2
Total numbers of data generated in the run	30000
Amount of sequences of 1000 data	30
Minimum time to generate 1000 data in the run	176 Seconds
Maximum time to generate 1000 data in the run	268 Seconds
Average time to generate 1000 data in the run	225.133 Seconds
Average time to generate 1000 data (first 12 sequences)	198.333 Seconds
Gradient – average increase of time every 1000 data in the run	3.005 Seconds
Gradient – average increase of time every 1000 data (first 12 sequences)	3.608 Seconds

4.2.3. DES model 2 – data generation time requirement

Due to higher complexity, the data generation time for model 2 is expected to be longer. To ensure that data can be obtained within a reasonable time frame, the number of replications per run has been reduced to 30,000. The time to generate data is aggregated for every 1000 replications, so there are 30 data points in a simulation run. The orange line in Figure 4 shows the data generation time of model 2. Similar to model 1, the time required to generate 1000 additional replications is not constant, but an increase of 3.005 seconds. The similarity can be seen through the two almost parallel lines shown in Figure 4. Table 5 summaries the time parameters of model 2.

4.3. DES model 3 – flexible manufacturing system

4.3.1. Simulation model 3 description

The third model is the most complex and adopted from Slack et al. (2010). Changes have been made to the model to better suit this research purpose. The manufacturing site produces four types of SKU, each has its own parameters including demand, weight, type, and processing time by station. The original production site contains both sequential and parallel processes, similar to the cases discussed above. However, in addition to the original scope, a flexible manufacturing facility has been introduced and included in the model. According to Das and Nagendra (1997), a flexible manufacturing facility can adapt to changes, both in its internal and external environment, quickly and economically; and particularly, to provide routing flexibility, which is the ability to manufacture a product via several alternate routes in the same facility. The flexible layout is set to

Table 6. Parameters of the SKUs in Model 3.

			Process time (sec)	
Part type	Weight (kg)	Cell capable of processing the SKU	Mean	Standard Deviation
SKU1	1.5	Cell 1	25	0.1
SKU2	2.5	Cell 1, Cell 2, Cell 3	15	0.1
SKU3	4	Cell 3, Cell 4	23	0.1
SKU4	4.5	Cell 1, Cell 2, Cell 4	17	0.1

become accepted by an increasing number of SME manufacturing and growing demand for customizable products as a part of e-manufacturing development stream, which also implies digitalization, mobility, and immediacy (Cheng & Bateman, 2008). Refer to Figure 2 for layout and RHC19 for detailed process flowcharts.

In this model, a forklift is used to move parts in batches from station to station. The forklift will wait until the total weight reaches 2,000 kilograms, which is the maximum load of the forklift. Each part is assigned a specific constant weight, corresponding to a specific SKU type. The SKU parameters are shown in Table 6. With the introduction of a forklift, according to the requirements of ARENA, the distance between stations needs to be added in the DES model. Moreover, travel between stations does take some stochastic time within a predefined range. In addition to the introduced travel time, the loading and unloading time at each station need to be modelled as a constant. If parts are queuing during the processing interval, it is assumed that there is an infinite buffer in front of the workstations.

The DES model starts when the raw material (coil) arrives at the blanking station, where it is processed into one of the four SKUs at the blanking station. Processing time includes setting time and blanking time. According to the SKU type, the blanks are then loaded and sent in batches by the forklift to the corresponding pressing stations, which work as independent parallel processes, for pressing. The post-pressing parts are then transferred to the assembly station in batches by the forklift. There are four flexible manufacturing cells in the assembly station. Each cell is capable of processing specific SKU types with different processing times and variances, and is equipped with an infinite input buffer, modelled as a warehouse with unlimited capacity, to hold incoming parts. The same transfer logic of using the forklift applies to moving the assembled parts to the last station for painting and quality checking. The paint and quality station mimics two conveyors, each capable of handling 3,600 parts at once, and a simple quality check station (Harun & Cheng, 2012). Processes including primer coating, painting, and furnace drying are performed in the cell, requiring a total process time of 90 minutes. Quality control is the last process, with processing time modelled as a triangular distribution. Immediately after the quality check, the SKU is considered as produced; and consolidated information on inter-arrival times, processing times, capacities, and other variables are recorded in the dataset.

4.3.2. DES model 3 – data features

Model 3 is the most complex, with series and parallel layouts and flexible workflows. The number of data features has increased significantly, and so is the time required to generate data. There are five groups of data features, and a total of 77 features, as shown in Table 7.

Table 7. Model 3 – matrix of data features.

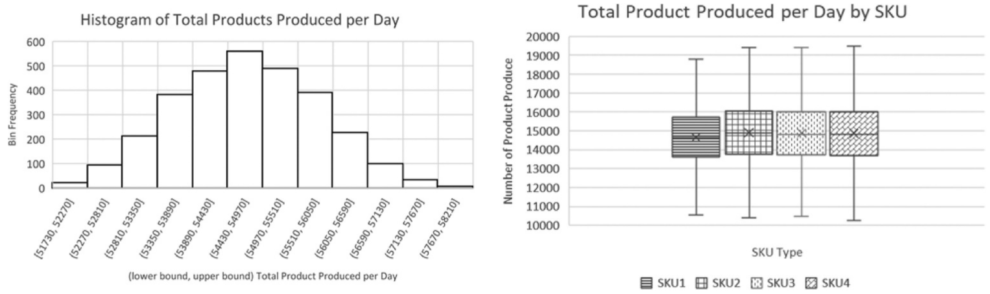
Group/Facility	Blanking	Pressing	Assembly	Paint & Quality	Forklift	Total Features
Utilisations	Blanking (1)[a]	Press 1–4 (4)	Cell 1–4 (4)	Paint 1–2 (2) Quality (1)	Forklift (1)	13
Queues	Uncoil SKU 1–4 (4) Blanking (1)	Press 1–4 (4)	Warehouse 1–4 (4) Cell 1–4 (4)	Paint 1–2 (2) Quality (1)	Blanking (1) Pressing (1) Assembly (1)	23
Part counters			Cell 1: SKU 1–4 (4) Cell 2: SKU 1–4 (4) Cell 3: SKU 1–4 (4) Cell 4: SKU 1–4 (4)	Product (1)		17
Cycle counters[b]			Cell 1: SKU 1–4 (4)			4
Time[c]	Value added time: SKU 1–4 (4) Waiting time: SKU 1–4 (4) Transporting time: SKU 1–4 (4) Other time: SKU 1–4 (4) Non-value added time: SKU 1–4 (4)					20

[a]The number of data features is shown in brackets. [b]Cycle counters are not production data, for validation purpose. [c]The accumulated time across facility.

The complexity of model 3 resembles a real production layout. Massive amount of data, a total of 730497 rows, was generated in this test study. In order to gain insight into the production, one of the runs containing 3000 replications was randomly selected for data visualisation. The distribution of total daily production, and the distribution of daily production by SKU, are shown in Figure 6. The total product assembled by SKU in each cell per day, and the distribution of assembly cell utilisation, are shown in Figure 7. These figures demonstrate that DES is capable of generating massive amount of well distributed synthetic data, covering a large number of production scenarios of a complex manufacturing layout.

4.3.3. DES model 3 – data generation time requirement

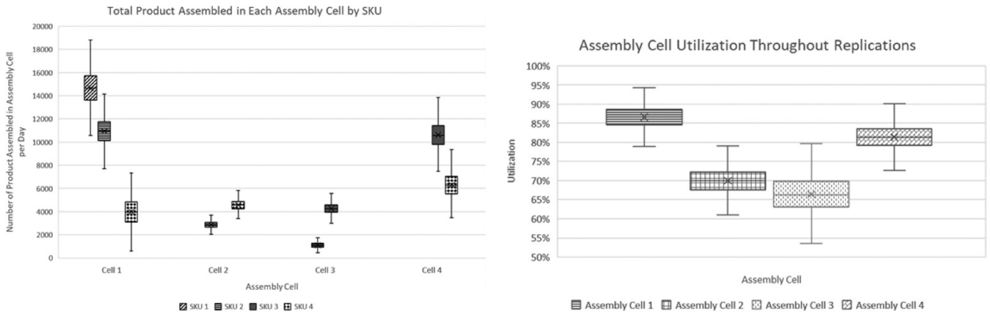
A dataset of 36000 replications was generated from 3 runs of 12,000 replications each. The time to generate data is aggregated for every 1000 replications, so there are 12 sequences of 1000 data in each run. The top three lines in Figure 4 show the data generation times of the three runs. The three lines show a similar upward trend of time required to generate every additional 1000 data points. Similar to models 1 and 2,



(a) Model 3 – histogram of total production per day.

(b) Model 3 – box plot of production by SKU per day.

Figure 6. Model 3 – total production and production by SKU per day.



(a) Model 3 – total product assembled by SKU in each cell per day.

(b) Model 3 – box plot of assembly cell utilsation.

Figure 7. Model 3 – total product assembled by SKU in each cell per day and assembly cell utilisation.

Table 8. Model 3 – summary of time requirement parameters for data generation.

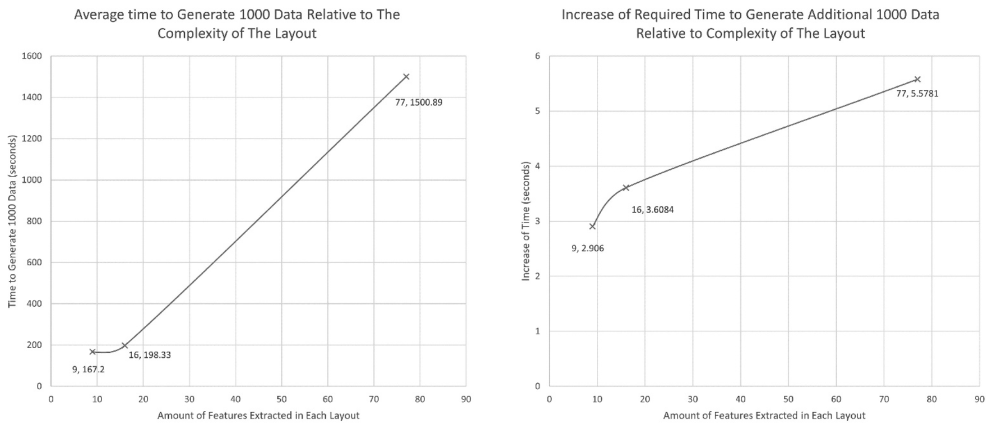
Parameter of Time Requirement	Run 1	Run 2	Run 3	Average
Total replication	12000	12000	12000	12000
Sequences of 1000 data	12	12	12	12
Minimum time to generate 1000 data (seconds)	1465	1464	1462	1463.67
Maximum time to generate 1000 data (seconds)	1590	1516	1504	1536.67
Average time to generate 1000 data (seconds)	1523	1492.25	1487.42	1500.89
Gradient – time increase of every 1000 data (seconds)	9.3427	4.3811	3.0105	5.5781

this phenomenon is suspected to be related to the gradual depletion of memory and other hardware or software related issues. A summary of time requirement parameters for data generation is provided in Table 8.

5. Discussion

The proposed framework has proven to be robust in synthetic data generation. Generally speaking, a more complex manufacturing system requires more data features to fully describe its system behaviour. Therefore, the number of data features can be regarded as a proxy for system complexity. Figure 8(a) shows that as the number of features (or model complexity) increases, the time required to generate data also increases. Due to the high complexity of the third model, the time required to generate synthetic data has increased by nearly 800% and 570% compared to the first and second layouts. The increase is significant, but not unexpected.

As mentioned before, the data generation speed tends to slow down gradually during a run, most likely due to memory and other hardware/software issues. Figure 8(b) shows that this extra time requirement, for generating 1000 additional data points in the same run, will also increase when the number of features (or model complexity) increases. Since the data generation time is specific to the model and computing environment, trial and error is necessary especially in the early stages to gain some insights for planning the entire data generation process. Another consideration is to find a practical balance between number of runs and number of replications to obtain the right amount of data.



(a) Average data generation time vs. data features.

(b) Increase in time vs. data features.

Figure 8. The average and marginal time increase vs. number of data features.

6. Limitations and future research

It has been demonstrated that using DES to generate synthetic data is a fast and reliable approach. This method is a viable alternative when real manufacturing data is not available or difficult to collect. An important limitation is that DES can be used to generate data for a steady-state production; while extreme/rare events, such as extended machine breakdowns or major supply chain interruptions, are not considered. In order to cover both steady-state production scenarios and extreme/rare events, different datasets with different data features may be required. It is also possible to add more features in the dataset to achieve wider coverage and solve more problems. For example, adding cost-related data features to the datasets will allow for investigation of production cost optimisation problems.

There are five research directions identified for future work:

(1) Synthetic data generation is useful for many different types of manufacturing problems, from analysis, planning, optimisation, to decision making. Conducting a meta-analysis of the classification and requirements of data features in different manufacturing problems will help confirm the applicability of DES to a much wider range of manufacturing problems, and provide valuable insights for further improvement of the proposed synthetic data generation framework.

(2) Investigating the resulting ML model performance by using real data and synthetic data. Real data collected naturally contains noise due to the environment and many other external and internal factors. Therefore, real data is never perfect and often suffers from the corruption that may hinder the performance of the ML algorithm (El Emam et al., 2020; Wu & Zhu, 2008). Moreover, there is a relationship between the real-world data and DES. When real data is used to construct the DES model and if the real data is corrupted, then the DES model will also be corrupted. To improve the ML model accuracy, data cleaning can be applied to identify the incorrect, incomplete, inaccurate, or missing data and

then modify, replace or delete accordingly. However, DES can produce synthetic data by reverse-engineering the real data to model its statistical properties and distributions. The problem with synthetic data is that generating good synthetic data is hard (Strickland, 2022). To extend the current research, designing and conducting experiments to analyse and compare the robustness, accuracy, and effectiveness of the ML models learned with cleaned real data, and synthetic data with different levels of controlled noise will help understand how to generate better synthetic data for manufacturing problems.

(3) Conducting a conceptual study on how real data and synthetic data, both as historical and real-time data, can be augmented by domain experts, from the perspectives of data, information, and knowledge. Designing and building frameworks that effectively integrate and enable human-machine collaboration in solving manufacturing problems.

(4) Conducting an experimental study in integrating real and synthetic data using existing tools to solve real-time manufacturing problems. Existing tools such as MTConnect and Apache NiFi are promising. MTConnect provides a standard solution to collect data from production machines and devices; while Apache NiFi supports automation of data flow between software systems (Cui et al., 2019). In addition to structured data, the experiments can also be extended to include unstructured data types that arrive into the systems as data-at-rest or data-in-motion (Cui et al., 2020b).

(5) DES and synthetic data generation should be considered as an integral part of the big data ecosystem (Cui et al., 2020a). The software tools (e.g. AI, ML, analytics, visualisation, and workflow) that currently exist in the ecosystems are powerful enablers for innovative manufacturing applications and solutions that bridge the virtual and real worlds. Applying these tools, together with DES and synthetic data, to develop creative applications is an exciting area of future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

K. C. Chan http://orcid.org/0000-0002-8756-2991

References

Allen, T. T. (2011). *Introduction to discrete event simulation and agent-based modeling: Voting systems, health care, military, and manufacturing*. Springer Science & Business Media.

Azadeh, A., Saberi, M., Kazem, A., Ebrahimipour, V., Nourmohammadzadeh, A., & Saberi, Z. (2013). A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization. *Applied Soft Computing*, *13*(3), 1478–1485. https://doi.org/10.1016/j.asoc.2012.06.020

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. *IADS-DM*.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481–485. https://doi.org/10.1257/aer.p20151021

Cheng, K., & Bateman, R. J. (2008). e-Manufacturing: Characteristics, applications and potentials. *Progress in Natural Science*, *18*(11), 1323–1328. https://doi.org/10.1016/j.pnsc.2008.03.027

Cui, Y., Kara, S., & Chan, K. C. (2019, November). Large scale MTConnect data collection. In *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)* (pp. 77–82). IEEE.

Cui, Y., Kara, S., & Chan, K. C. (2020a). Manufacturing big data ecosystem: A systematic literature review. *Robotics and computer-integrated Manufacturing*, *62*, 101861. https://doi.org/10.1016/j.rcim.2019.101861

Cui, Y., Kara, S., & Chan, K. C. (2020b, December). Monitoring and control of unstructured manufacturing big data. In *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 928–932). IEEE.

Das, S. K., & Nagendra, P. (1997). Selection of routes in a flexible manufacturing facility. *International Journal of Production Economics*, *48*(3), 237–247. https://doi.org/10.1016/S0925-5273(96)00106-5

Denkena, B., Schmidt, J., & Krüger, M. (2014). Data mining approach for knowledge-based process planning. *Procedia Technology*, *15*, 406–415. https://doi.org/10.1016/j.protcy.2014.09.095

El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation: Balancing privacy and the broad availability of data*. O'Reilly Media.

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data mining and analytics in the process industry: The role of machine learning. *IEEE Access*, *5*, 20590–20616. https://doi.org/10.1109/ACCESS.2017.2756872

Greasley, A., & Edwards, J. S. (2021). Enhancing discrete-event simulation with big data analytics: A review. *Journal of the Operational Research Society*, *72*(2), 247–267. https://doi.org/10.1080/01605682.2019.1678406

Gyulai, D., Kádár, B., & Monostori, L. (2014). Capacity planning and resource allocation in assembly systems consisting of dedicated and reconfigurable lines. *Procedia CIRP*, *25*, 185–191. https://doi.org/10.1016/j.procir.2014.10.028

Harun, K., & Cheng, K. (2012). An integrated modeling method for assessment of quality systems applied to aerospace manufacturing supply chains. *Journal of Intelligent Manufacturing*, *23*(4), 1365–1378. https://doi.org/10.1007/s10845-010-0447-7

Huang, H. H., Pei, W., Wu, H. H., & May, M. D. (2013). A research on problems of mixed-line production and the re-scheduling. *Robotics and Computer-Integrated Manufacturing*, *29*(3), 64–72. https://doi.org/10.1016/j.rcim.2012.04.014

Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., Van de Walle, R., & Van Hoecke, S. (2016). Convolutional neural network based fault detection for rotating machinery. *Journal of Sound and Vibration*, *377*, 331–345. https://doi.org/10.1016/j.jsv.2016.05.027

Kaylani, H., & Atieh, A. M. (2016). Simulation approach to enhance production scheduling procedures at a pharmaceutical company with large product mix. *Procedia CIRP*, *41*, 411–416. https://doi.org/10.1016/j.procir.2015.12.072

Koh, S. C. L., & Saad, S. M. (2003). MRP-controlled manufacturing environment disturbed by uncertainty. *Robotics and Computer-Integrated Manufacturing*, *19*(1–2), 157–171. https://doi.org/10.1016/S0736-5845(02)00073-X

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, *51*(11), 1016–1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Lechevalier, D., Shin, S. J., Woo, J., Rachuri, S., & Foufou, S. (2015). A virtual milling machine model to generate machine-monitoring data for predictive analytics. In *IFIP International Conference on Product Lifecycle Management* (pp. 835–845). Springer.

Lee, H., Kim, S. G., Park, H. W., & Kang, P. (2014). Pre-launch new product demand forecasting using the bass model: A statistical and machine learning-based approach. *Technological Forecasting and Social Change*, *86*, 49–64. https://doi.org/10.1016/j.techfore.2013.08.020

Maas, S. L., & Standridge, C. R. (2005, December). Applying simulation to interative manufacturing cell design. In *Proceedings of the Winter Simulation Conference*, IEEE.

Nyemba, W. R., & Mbohwa, C. (2017). Modelling, simulation and optimization of the materials flow of a multi-product assembling plant. *Procedia Manufacturing*, *8*, 59–66. https://doi.org/10.1016/j.promfg.2017.02.007

Pfeiffer, A., Gyulai, D., Kádár, B., & Monostori, L. (2016). Manufacturing lead time estimation with the combination of simulation and statistical learning methods. *Procedia CIRP*, *41*, 75–80. https://doi.org/10.1016/j.procir.2015.12.018

Priore, P., de la Fuente, D., Puente, J., & Parreño, J. (2006). A comparison of machine-learning algorithms for dynamic scheduling of flexible manufacturing systems. *Engineering Applications of Artificial Intelligence*, *19*(3), 247–255. https://doi.org/10.1016/j.engappai.2005.09.009

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Shahzad, A., & Mebarki, N. (2012). Data mining based job dispatching using hybrid simulation-optimization approach for shop scheduling problem. *Engineering Applications of Artificial Intelligence*, *25*(6), 1173–1181. https://doi.org/10.1016/j.engappai.2012.04.001

Shiue, Y. R. (2009). Data-mining-based dynamic dispatching rule selection mechanism for shop floor control systems using a support vector machine approach. *International Journal of Production Research*, *47*(13), 3669–3690. https://doi.org/10.1080/00207540701846236

Silva, N., Ferreira, L. M. D., Silva, C., Magalhães, V., & Neto, P. (2017). Improving supply chain visibility with artificial neural networks. *Procedia Manufacturing*, *11*, 2083–2090. https://doi.org/10.1016/j.promfg.2017.07.329

Slack, N., Chambers, S., & Johnston, R. (2010). *Operations management*. Pearson education.

Strickland, E. (2022, February 17). *Are you still using real data to train your AI? IEEE Spectrum*. https://spectrum.ieee.org/synthetic-data-ai

Subramaniyan, M., Skoogh, A., Muhammad, A. S., Bokrantz, J., Johansson, B., & Roser, C. (2020). A generic hierarchical clustering approach for detecting bottlenecks in manufacturing. *Journal of Manufacturing Systems*, *55*, 143–158. https://doi.org/10.1016/j.jmsy.2020.02.011

Subramaniyan, M., Skoogh, A., Salomonsson, H., Bangalore, P., Gopalakrishnan, M., & Sheikh Muhammad, A. (2018). Data-driven algorithm for throughput bottleneck analysis of production systems. *Production & Manufacturing Research*, *6*(1), 225–246. https://doi.org/10.1080/21693277.2018.1496491

Van der Zee, D. J., & Van der Vorst, J. G. (2007, December). Guiding principles for conceptual model creation in manufacturing simulation. In *2007 Winter Simulation Conference*, (pp. 776–784). IEEE.

Weichert, D., Link, P., Stoll, A., Rüping, S., Ihlenfeldt, S., & Wrobel, S. (2019). A review of machine learning for the optimization of production processes. *The International Journal of Advanced Manufacturing Technology*, *104*(5), 1889–1902. https://doi.org/10.1007/s00170-019-03988-5

Weimer, D., Scholz-Reiter, B., & Shpitalni, M. (2016). Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *CIRP Annals*, *65*(1), 417–420. https://doi.org/10.1016/j.cirp.2016.04.072

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, (Vol. 1). Springer-Verlag.

Wu, X., & Zhu, X. (2008). Mining with noise knowledge: Error-aware data mining. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *38*(4), 917–932. https://doi.org/10.1109/TSMCA.2008.923034

Wuest, T., Irgens, C., & Thoben, K. D. (2014). An approach to monitoring quality in manufacturing using supervised machine learning on product state data. *Journal of Intelligent Manufacturing*, *25*(5), 1167–1180. https://doi.org/10.1007/s10845-013-0761-y

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. *Production & Manufacturing Research*, *4*(1), 23–45. https://doi.org/10.1080/21693277.2016.1192517

Zhang, Y., Ma, S., Yang, H., Lv, J., & Liu, Y. (2018). A big data driven analytical framework for energy-intensive manufacturing industries. *Journal of Cleaner Production*, *197*, 57–72. https://doi.org/10.1016/j.jclepro.2018.06.170

Zhuo, L., Chua Kim Huat, D., & Wee, K. H. (2012). Scheduling dynamic block assembly in shipbuilding through hybrid simulation and spatial optimisation. *International Journal of Production Research*, *50*(20), 5986–6004. https://doi.org/10.1080/00207543.2011.639816