Decision Analytics Journal 19 (2026) 100721

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A statistical analytics framework for decision-ready scheduling in autonomous intralogistics

Ricardo Moura [a,b,*], Nuno Pessanha Santos [b,c,d], Ana Madureira [e], Alexandra Tenera [f]

[a] Center for Mathematics and Applications (NOVA Math), Campus de Caparica, 2829-516, Almada, Portugal
[b] Portuguese Navy Research Center (CINAV), Portuguese Naval Academy, Base Naval de Lisboa, 2810-001, Almada, Portugal
[c] Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), Avenida Rovisco Pais 1, 1049-001, Lisbon, Portugal
[d] Portuguese Military Research Center (CINAMIL), Portuguese Military Academy, Avenida Conde Castro Guimarães, 2720-113, Amadora, Portugal
[e] Imeguisa Portugal - Indústrias Metálicas Reunidas S.A., Rua 5 de Outubro s/n, Apartado 57, 2950-727, Palmela, Portugal
[f] UNIDEMI Research Center, NOVA University of Lisbon, Campus de Caparica, 2829-516, Almada, Portugal

ARTICLE INFO

Keywords:
Intralogistics analytics
Operational efficiency
Robust scheduling
Statistical distribution fitting
Performance analysis
Discrete events

ABSTRACT

The widespread adoption of Internet of Things (IoT) technologies within Industry 4.0 has increased the need for more responsive and efficient intralogistics operations, especially as industries move toward highly customized, small-batch production. Autonomous Mobile Robots (AMRs) provide the flexibility needed for dynamic industrial environments. However, their effective integration relies on accurate task-duration estimates, which are difficult to obtain due to system variability and operational uncertainty. This work addresses this challenge by statistically modeling picking and loading/unloading times using real-world industrial data. Probability distribution-fitting techniques are employed to capture the random nature of these operations, thereby reducing the need for unrealistic assumptions often used in scheduling models. The proposed methodology enhances the statistical basis for robust AMR fleet scheduling, improving coordination and reducing operational variability. Validation is conducted by directly comparing empirical data with fitted distributions, demonstrating clear improvements in representativeness and predictive accuracy. The results demonstrate the potential of this approach to improve decision-making in highly dynamic intralogistics environments.

1. Introduction

The 4th Industrial Revolution has driven the adoption of Internet of Things (IoT) technologies, enhancing business competitiveness and globalizing Industry 4.0 technologies [1]. This transformation integrates production engineering, information and communication systems, and big data analytics to shift from mass production to small-batch, highly customized manufacturing. However, the increasing market demand for product variety and customization requires greater responsiveness from industries and their collaborative networks [2].

To maintain high productivity, companies must either invest in new production lines or improve existing ones, while minimizing opportunity costs through optimized intralogistics operations [3]. The efficiency of intralogistics operations can be enhanced by utilizing Autonomous Mobile Robots (AMRs), which can quickly adapt to the dynamic environments of industrial settings. However, their use introduces new integration challenges, such as developing approaches to route planning that account for unexpected events, including breakdowns and uncertainty in task execution times [4].

A critical step toward effective AMR use is task scheduling, which requires reliable estimates of task duration. This is challenging given the dynamic environment in which they operate (e.g., moving obstacles). Accurate task duration estimates reduce system variability and improve fleet coordination, just as they improve project scheduling under uncertain durations [5]. Precise modeling of picking times is essential for planning future scheduling approaches, including robust scheduling methods. While Mixed-Integer Linear Programming (MILP) is often used for task allocation and scheduling, its success depends on accurate time distributions and predictive modeling [6,7]. In line with optimization-driven research on robotic manufacturing systems [8], the practical effectiveness of AMR scheduling models ultimately depends on realistic input parameterization, motivating distribution-based characterization of uncertain service times to enable decision-ready and robust scheduling.

Implementing AMR fleets or other complex systems in real-world settings benefits from a digital representation of the problem, ranging from a single simulator to a complete Digital Twin, which allows

* Corresponding author at: Portuguese Navy Research Center (CINAV), Portuguese Naval Academy, Base Naval de Lisboa, 2810-001, Almada, Portugal.
E-mail address: rp.moura@fct.unl.pt (R. Moura).

https://doi.org/10.1016/j.dajour.2026.100721
Received 12 January 2026; Received in revised form 27 April 2026; Accepted 25 May 2026
Available online 28 May 2026

testing algorithms and proposed models under actual deployment conditions [1,9,10]. A range of methods for incorporating input data into simulation models has been proposed and explored, including trace-driven simulation, empirical distributions, and statistical inference [11]. Discrete Event Simulation (DES) enables modeling of systems that change at discrete points in time [11], supporting model verification and validation and ensuring that Key Performance Indicators (KPIs) are met. It can create physical or mathematical models of varying complexity, based on available resources and decision-making needs. As a stochastic, dynamic, and discrete approach, DES can model uncertainty, time-dependent behaviors, and changes in discrete states, thereby ensuring a realistic representation of the system. However, a significant limitation of DES is its reliance on accurate models of the target system to provide acceptable accuracy [10,12].

Real-world systems are inherently subject to randomness that affects model validity and operational performance, making realistic input probability distributions essential. The operational data used in this study were collected and preprocessed in the context of an industrial case study developed under Project AGiLE [13–15]. This work improves the robustness of AMR scheduling models by accurately modeling the statistical behavior of picking and loading/unloading operations. Using real-world data, probability distribution-fitting techniques are applied to capture representative statistical characteristics and reduce reliance on unrealistic assumptions. These improvements provide a stronger statistical foundation for task allocation, ensuring more reliable and adaptable scheduling by enabling accurate timing information and reducing bottlenecks. The proposed approach is validated by directly comparing the empirical distribution with the newly proposed distributions on processed data, demonstrating its potential to enhance decision-making in dynamic operational environments.

The remainder of this article is organized as follows. Section 2 reviews the state of the art on AMRs as intralogistics agents and on robust scheduling, highlighting the main challenges reported in the literature. Section 3 formulates the scheduling problem under study and states the modeling assumptions. Section 4 details the proposed methodology and the statistical modeling framework designed to better reflect real operational conditions. Section 5 presents the experimental evaluation and performance analysis. Section 6 describes how to generate synthetic data with statistical properties consistent with the original recorded samples for further testing and analysis. Finally, Section 7 concludes the article and outlines directions for future research.

2. Related work

This section reviews the existing state of the art on AMRs as intralogistics agents (Section 2.1), on robust scheduling (Section 2.2), and on the statistical modeling of operational times for data-driven decision support (Section 2.3).

2.1. Autonomous mobile robots as intralogistics agents

AMRs autonomously transport materials in industrial environments, supporting intralogistics by resupplying production lines and handling flows from raw materials to finished products [10]. Such vehicles can be guided using laser-based systems, wireless communication sensors, barcode markers, Radio Frequency Identification (RFID) technology, magnetic guidance strips, optical navigation systems, ultrasonic or vision-based systems, and Global Navigation Satellite Systems (GNSS) [3,16,17].

An essential feature of AMR operation is route planning, which depends on vehicle localization and onboard sensors to interact with the surrounding environment and gather relevant data [1,9]. Different approaches have been explored, such as knowledge-guided multi-objective Estimation of Distribution Algorithms with delivery satisfaction evaluation [18], Hybrid Genetic Algorithms combined with Large Neighborhood Search (GA-LNS) [19], Digital Twin-based solutions [10], and simulators for intralogistics automation based on DES and agent-based modeling [1,9]. More broadly, integrated decision-analytic approaches have shown that DES can be combined with complementary system-analysis methods to support structured diagnosis and improvement of Industry 4.0 manufacturing systems [20], reinforcing the need for credible, data-grounded input models in simulation-based decision support.

Monitoring AMR fleet performance is crucial for enhancing intralogistics efficiency and productivity. KPIs such as material replenishment time, vehicle punctuality, and unloading efficiency depend on machine stoppages, last-minute orders, or operator unavailability [18]. Additional metrics, including task completion rate and incident frequency, highlight the need for improved route planning and environmental adaptation to AMR traffic [1,18].

Vehicle scheduling and route planning remain significant challenges in intralogistics, as they directly affect material delivery efficiency and overall productivity [18]. These tasks must also account for the limited energy autonomy of battery-powered AMRs, necessitating the optimization of routes and charging schedules [19]. Safety is another key concern, demanding proper speed control, braking distances, and dedicated human–robot interaction zones to protect personnel and equipment [10]. Furthermore, the introduction of AMRs may raise concerns among operators, making effective communication and change management essential for user acceptance [21].

2.2. Robust scheduling

Modern intralogistics must evolve towards Cyber-Physical Systems that are flexible, robust to unforeseen events, rapidly adaptable to new processes, and capable of continuously collecting information about the system's current state and operational needs. In this context, Robust Scheduling has gained increasing relevance, as it aims to construct operational plans for a set of entities while accounting for dynamic and unpredictable events that may interfere with their execution [22].

These scheduling problems are typically divided into two main questions: finding an optimal scheduling solution and evaluating its performance under disturbances. More recently, a new perspective has emerged that integrates environmental sustainability into scheduling problems, seeking solutions that not only minimize operational costs and the use of scarce resources but also reduce environmental impact [23].

To address such challenges, a wide range of mathematical formulations and algorithmic strategies have been proposed, supported by solvers that identify the best solution among multiple feasible alternatives. Among these, Mixed-Integer Linear Programming (MILP) remains one of the most widely adopted approaches due to its balance between modeling flexibility and computational tractability. Other notable methods include the Critical Chain Scheduling [24–26], Genetic Algorithms (GA) [27,28], Neural Networks (NNs) [29], Mixed-Integer Nonlinear Programming (MINLP) [30], Dynamic Event-Response Strategies (DERS) [31], the Analytic Hierarchy Process (AHP) as a multi-criteria decision-making framework [32], and the Non-dominated Sorting Genetic Algorithm II (NSGA-II), a well-established multi-objective optimization technique [23].

Despite of the advances, recent studies have predominantly focused on production and maintenance planning, while comparatively limited attention has been devoted to the robust scheduling of AMRs in industrial environments. This limitation highlights the need for scheduling approaches that are better aligned with the uncertainty and variability of real intralogistics operations.

2.3. Statistical modeling of operational times

A complementary research stream concerns the statistical modeling of operational times used as stochastic inputs to simulation, optimiza-

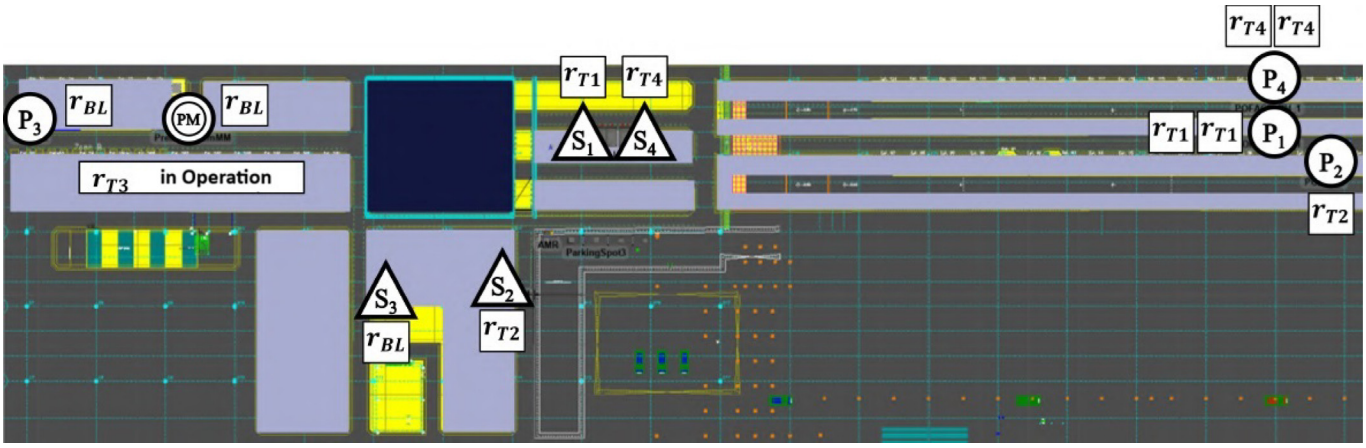

Fig. 1. Final assembly line layout highlighting the Manufacturing Supermarket (MS) (triangle markers) and the Point of Fit (PoF) (circle markers). The gray and dark blue areas represent zones with restricted AMR access.

tion, and scheduling models. In real industrial systems, processing and service times are inherently variable, making the specification of realistic input probability distributions a critical step in model development. In this context, the literature on input uncertainty has shown that neglecting the statistical uncertainty associated with input models may lead to biased performance estimates and less reliable decision support [33].

Within warehousing and intralogistics, several studies have emphasized the importance of modeling full probability distributions of operational times rather than relying solely on average values. For example, throughput-time distribution analysis has been used to support operational decision-making in warehouse systems [34], while more recent studies have examined order-picking times using real operational data and analytics-based estimation models [35,36]. Collectively, these works show that operational durations are heterogeneous and context-dependent, thereby supporting the use of empirical and fitted probability distributions instead of simplified deterministic assumptions.

This perspective is also increasingly reflected in digital-twin and manufacturing applications. Recent studies have calibrated intralogistics digital twins using real data from plant sensors and information systems [37], and live-fitting approaches have been proposed to continuously identify probability distributions for process times from real-world production data [38]. Such developments reinforce the view that statistically grounded temporal models are essential when the objective is to improve operational robustness under changing and uncertain conditions.

Despite these advances, limited attention has been devoted to explicitly characterizing picking and loading/unloading times as stochastic inputs to AMR scheduling models in industrial environments. In particular, the literature still lacks studies that integrate real-world operational data, distribution fitting, and scheduling-oriented validation within a unified framework. This gap motivates the present work, which seeks to enhance the robustness of AMR scheduling by improving the statistical realism of the operational times on which task-allocation decisions depend.

3. Problem formulation

The system architecture, inspired by the AGiLE project [13–15], adopts a three-tier structure comprising a fleet of three AMRs, sensorized peripherals, and a supervisory entity. The system manages component transport between the Manufacturing Supermarket (MS) and the Point of Fit (PoF). The MS serves as an intermediate storage area for components or work-in-progress items, temporarily storing them to supply downstream processes. At the same time, the PoF corresponds to the specific location along the assembly line where components are integrated into the final product. Sequenced components share a unique code with their corresponding final product. They are delivered to the assembly point either from the MS during the picking process or pre-sequenced by the supplier.

The AMRs operate autonomously within the factory layout, navigating freely while avoiding both static and dynamic obstacles and communicating seamlessly with the peripherals (racks) and the supervisory entity. Two types of racks are defined in the system: Line-side Racks (r_{BL}), which are fixed structures located along the assembly line and typically used for non-sequenced components, and Transport Racks (r_{T}), which are mobile structures that carry sequenced components from the MS to the PoF and function as line-side racks when positioned at the PoF, as represented in Fig. 1.

The supply of components comprises three main processes [14,15], as illustrated in Fig. 2. The LH-WA component family uses three transport racks, each capable of carrying 24 pieces. The CF process employs two transport racks, each with a capacity of 18 pieces. The EM process operates with a single transport rack containing two boxes, each with 12 pieces. In this case, loading and unloading are performed via a shooter mechanism that transfers full or empty boxes between the line-side racks and the MS, both of which are fixed in position. A fourth process, RH-WA, serves as a backup and expansion point and uses the same rack capacity as LH-WA. This configuration enables coordination among multiple MSs and PoFs, ensuring a systematic, continuous supply of components in response to event-based triggers.

Each component is associated with a specific MS–PoF pair, as illustrated in Fig. 1, which defines four main processes labeled 1, 2, 3, and 4. In this representation, triangles indicate the locations of the MSs (S_1, S_2, S_3, and S_4), circles mark the PoFs (P_1, P_2, P_3, and P_4), and double circles denote the Pre-Assembly (PA) process for the EM, which occurs before its assembly at PoF P_3. The supply of these components involves bidirectional rack movements between each MS and its corresponding PoF. These transfers are triggered by production events, typically task initiation, process completion, or low inventory detection. The WA-RH, when active, operates under the same triggers as the WA-LH, since both correspond to the same component type installed on opposite sides of the line. Once a trigger is activated, the supervisory system generates a transport request, which the AMR fleet management module evaluates to assign the nearest available robot to execute the task.

The rack movement logic follows a configuration defined by the project managers, in which the combination of the number of pieces *W*, *X*, *Y*, and *Z*, as shown in Table 1, determines the triggering of transport operations. This configuration was validated in a previously performed study on an assembly line [14]. The timing of these triggers depends on several factors, including the picking time of each component at its

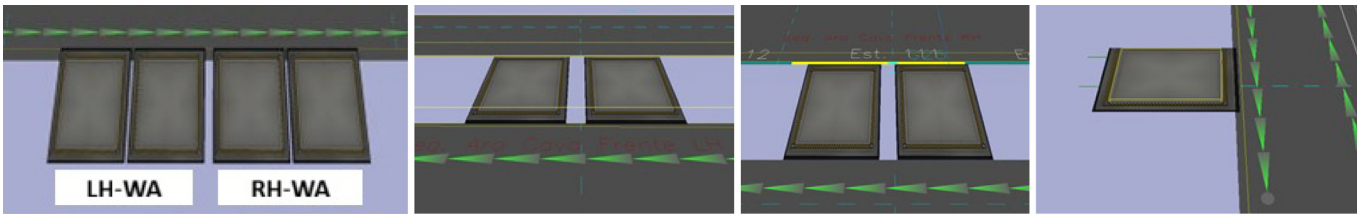

Fig. 2. Different queues: MS for LH-WA and RH-WA (*left*), PoF for LH-WA (*center left*), PoF for RH-WA (*center right*), and MS for CF (*right*).

Table 1
Trigger events for rack movement by component and location.

	LH-WA	CF	EM
MS	Rack exits PoF (free PoF position) Picking completed	Picking completed Rack exits PoF (free PoF position)	W=12 pieces needed at pre-assembly site
PoF	X=4 pieces in rack being consumed at PoF	Y=4 pieces in rack being consumed at PoF	Box unloading completed Box loading completed
PA	–	–	Z=12 pieces needed at PoF

MS, the loading and unloading durations at the start and end of each supply task, AMR travel speeds, box handling times for the EM, and the number of remaining pieces at the assembly line. In the case of the CF process, one of the triggers in Table 1 is redundant, as two rack positions are always available at the PoF, ensuring that a free position is maintained whenever a rack is located at the MS.

The loading and unloading time of an AMR corresponds to the duration between its arrival at a loading or unloading area, the handling of the rack, and the moment it resumes movement. There are several different queue spaces, each serving as a waiting area where entities are temporarily held before moving on to the next stage of the system, as illustrated in Fig. 2.

To access real data for the times of every task, information was collected from four shifts (Shift 1–4) at Volkswagen Autoeuropa, each with distinct, non-overlapping operators [13]. To model the plant under optimal performance, the picking times for each shift were gathered during April and May 2022, followed by an analysis to ensure their consistency and validity for subsequent development [14]. The recorded picking times, which indicate how long it takes to complete specific tasks, provide a reliable basis for setting simulation model parameters. All observations with picking times of less than one minute were excluded from the dataset. This threshold was established to mitigate the impact of time recording inconsistencies, which frequently resulted in unrealistically short durations that did not accurately reflect the execution of picking tasks. Without validated reference times, this filtering step was used as a cautious data-cleaning measure.

4. Statistical modeling of picking time distributions

Real-world systems inherently exhibit randomness that affects the validity of models. The quality and type of input data, such as trace-driven simulation, empirical distributions, or fitted probability distributions, influence model complexity [11]. In the described approach, we focus on adjusting probability distributions that best fit the AMR picking and load/unload times. All implementations were conducted in Python, using SciPy for statistical modeling and the Fitter library for distribution fitting [39]. To ensure that the most appropriate input distributions are selected for the model, the picking times, together with the AMR load/unload time data, are preprocessed through the following steps:

1. **Outliers Removal (**Section 4.1**)** – Observations with picking times below 60 s were excluded as implausibly short records. Any additional screening was restricted to the upper tail, using a skewness-aware upper fence for clearly right-skewed samples and the standard Tukey upper fence otherwise;
2. **Non-Parametric Density Estimation (**Section 4.2**)** – A non-parametric estimation of the probability density function is obtained to characterize the underlying data distribution;
3. **Distribution Fitting (**Section 4.3**)** – Several statistical distributions are fitted to the data to identify suitable parametric models;
4. **Performance Metrics Analysis (**Section 4.4**)** – The goodness-of-fit is assessed to evaluate the fitting quality and determine the most representative model.

4.1. Outliers removal

To ensure data consistency while preserving operationally relevant delay events, preprocessing was conducted in two stages. First, observations with picking times below 60 s were removed as implausibly short records. This threshold was introduced as a data-quality rule to reduce the influence of time-recording inconsistencies that generated unrealistically short durations.

After this initial filtering step, no further trimming was applied to the lower tail, and any additional outlier screening was restricted to the upper tail only. This choice reflects the intended scheduling application, in which unusually long task durations are more critical than unusually short ones.

For each sample, let the InterQuartile Range (IQR) be defined as $Q_3 - Q_1$, where Q_1 and Q_3 denote the 25th and 75th percentiles, respectively. We then computed the standard Tukey upper fence [40] as

$$\mathrm{UB}_{\mathrm{Tukey}} = Q_3 + 1.5 \times \mathrm{IQR}. \tag{1}$$

To account for skewness, we also evaluated the robust medcouple statistic (MC) [41]. For positively skewed samples, the upper fence was adjusted using the medcouple-based rule of the adjusted boxplot for skewed distributions [42]. Specifically, when $MC > 0.10$, the upper bound was defined as

$$\mathrm{UB}_{\mathrm{adj}} = Q_3 + 1.5\, e^{3MC} \times \mathrm{IQR}, \tag{2}$$

whereas, for $MC \leq 0.10$, the standard Tukey upper fence was retained. Thus, the final upper screening threshold was

$$\mathrm{UB} = \begin{cases} Q_3 + 1.5\, e^{3MC} \times \mathrm{IQR}, & \text{if } MC > 0.10, \\ Q_3 + 1.5 \times \mathrm{IQR}, & \text{otherwise.} \end{cases} \tag{3}$$

Table 2
Probability density functions and parameter definitions for selected continuous distributions [43–47].

Distribution	Probability density function	Parameters
Generalized Normal	$f(x;\mu,\alpha,\beta) = \frac{\beta}{2\alpha\,\Gamma(1/\beta)} \exp\left[-\left(\frac{\lvert x-\mu\rvert}{\alpha}\right)^{\beta}\right]$	μ: location, $\alpha > 0$: scale, $\beta > 0$: shape
Skew Normal	$f(x;\xi,\omega,\alpha) = \frac{2}{\omega}\,\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\,\frac{x-\xi}{\omega}\right)$	ξ: location, $\omega > 0$: scale, α: shape
Beta	$f(x;\alpha,\beta,\mu,\sigma) = \frac{1}{\sigma\,B(\alpha,\beta)}\left(\frac{x-\mu}{\sigma}\right)^{\alpha-1}\left(1-\frac{x-\mu}{\sigma}\right)^{\beta-1}$	$\alpha > 0$: shape, $\beta > 0$: shape, μ: location, $\sigma > 0$: scale, $\mu < x < \mu+\sigma$
Generalized Logistic	$f(x;\mu,\sigma,c) = \frac{c}{\sigma}\,\frac{\exp\left[-(x-\mu)/\sigma\right]}{\left(1+\exp\left[-(x-\mu)/\sigma\right]\right)^{c+1}}$	μ: location, $\sigma > 0$: scale, $c > 0$: shape

Only observations above UB were removed at this stage. This asymmetric preprocessing strategy limits excessive trimming of plausible upper-tail durations in clearly right-skewed samples while preserving a conservative screening rule for approximately symmetric distributions.

4.2. Non-parametric density estimation

Kernel Density Estimation (KDE) [48] provides a nonparametric estimate of the underlying probability density function, offering a smooth representation of the data distribution while reducing the influence of sampling noise [49]. It is defined as

$$\hat{f}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_n}\right), \tag{4}$$

where n is the sample size, x is the evaluation point, x_i is the ith observed value in the sample, and h_n is the bandwidth (or smoothing parameter), which controls the smoothness of the estimate, similarly to the bin width in histograms. The kernel function $K(\cdot)$ defines the smoothing shape. The default kernel is the Gaussian distribution, meaning that for each x_i, a Gaussian probability density function is centered at $\mu = x_i$ with standard deviation $\sigma = h_n$. The weighted sum of these n Gaussian kernels, scaled by $\frac{1}{nh_n}$, yields a continuous estimate of the underlying probability density.

4.3. Distribution fitting

The variability in task durations was initially modeled using Gaussian Mixture Models (GMMs), which provide a flexible probabilistic framework for capturing multimodal behavior. Each sample $\mathbf{X} = \{x_1,\ldots,x_n\}$ is represented as a weighted combination of N Gaussian components, assuming latent subpopulations characterized by mean μ_j, variance σ_j^2, and non-negative mixture coefficient π_j according to

$$\sum_{j=1}^{N}\pi_j = 1, \qquad \pi_j \geq 0. \tag{5}$$

The PDF of the mixture is defined as

$$f(x_i \mid \theta) = \sum_{j=1}^{N}\pi_j f_j(x_i \mid \mu_j,\sigma_j^2), \tag{6}$$

with each component given by

$$f_j(x_i \mid \mu_j,\sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right). \tag{7}$$

The full parameter vector is represented as

$$\theta = \{\pi_1,\mu_1,\sigma_1^2,\ldots,\pi_N,\mu_N,\sigma_N^2\}. \tag{8}$$

Parameter estimation was performed using the Expectation–Maximization (EM) algorithm, which iteratively maximizes the log-likelihood represented by

$$\ell(\theta) = \sum_{i=1}^{n}\log f(x_i \mid \theta). \tag{9}$$

This formulation enables the identification of latent subgroups within the data, reflecting variations in operator, shift, or working condition. Compared with unimodal distributions (e.g., Normal, Log-normal, Weibull), GMMs capture heterogeneity and provide a more realistic statistical description of the system's temporal behavior, forming the probabilistic basis for the subsequent scheduling model.

In this study, GMMs were fitted to the picking-time distributions of operators at the MS stations and to AMR load/unload times. This allowed the detection of potential differences across shifts or working teams. Nevertheless, following Occam's Razor, simpler parametric models were preferred whenever they provided comparable goodness-of-fit. Thus, each sample was also fitted to a broad set of classical distributions with the Fitter library [39]. The representative families retained for reporting and discussion in this study are the Generalized Normal, Skew Normal, Beta, and Generalized Logistic distributions, as summarized in Table 2.

4.4. Performance metrics analysis

To assess the adequacy of the fitted probability models, we employed both the K-Sample Anderson–Darling (AD) [50] and the K-Sample Kolmogorov–Smirnov (KS) [51] tests. In the K-sample setting, the AD statistic assigns greater weight to tail discrepancies across all groups, making it particularly effective at detecting differences in extreme quantiles. In contrast, the K-sample KS test is most sensitive to differences around the central region of the pooled distribution. The complementary sensitivity profiles of the K-sample AD and KS procedures, therefore, provide a more comprehensive evaluation of goodness-of-fit across the entire domain. Both tests are nonparametric methods that assess the discrepancy between the empirical sample and simulations from the fitted models.

In this implementation, we use the two-sample AD test statistic to compare two independent samples $\{X_1,\ldots,X_n\}$ and $\{Y_1,\ldots,Y_m\}$ by quantifying the discrepancy between their empirical distribution functions. Let $N_{Tot} = n+m$ and let $\{Z_1,\ldots,Z_{N_{Tot}}\}$ denote the pooled ordered sample. For each $i = 1,\ldots,N_{Tot}$, the joint empirical distribution is defined as

$$H_i = \frac{i}{N_{Tot}}. \tag{10}$$

The empirical cumulative distribution functions of the individual samples, evaluated at Z_i, are given by

$$F_i = \frac{1}{n}\sum_{j=1}^{i}\mathbf{1}\{Z_j \in X\}, \tag{11}$$

and

$$G_i = \frac{1}{m}\sum_{j=1}^{i} \mathbf{1}\{Z_j \in Y\}. \tag{12}$$

The two-sample Anderson–Darling statistic is then computed as

$$A^2 = \frac{nm}{N_{Tot}} \sum_{i=1}^{N_{Tot}} \frac{\left(F_i - G_i\right)^2}{H_i\left(1 - H_i\right)}, \tag{13}$$

where large values of A^2 indicate increasing discrepancy between the two samples, particularly because the weighting term $\left[H_i(1-H_i)\right]^{-1}$, which assigns greater influence to differences in the tails. Because the model parameters are estimated from the same observed sample used for evaluation, standard calibrated p-values are not formally valid in this setting. To account for parameter estimation, the significance of the observed AD statistic was therefore assessed using a parametric bootstrap procedure, which provides an empirical reference distribution for the test statistic under the fitted model. Bootstrap-based p-values were computed using $B = 499$ replicates. In each replicate, a bootstrap sample of size n was generated from the fitted model, the same model family was refitted to that bootstrap sample, and the AD and KS statistics were recomputed by comparing the bootstrap sample with a new synthetic sample generated from the refitted model. The bootstrap p-value was then estimated as

$$\hat{p} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_b \geq T_{\text{obs}}\}}{B+1}, \tag{14}$$

where T denotes either the AD or KS statistic. In contrast, the two-sample KS test directly compares the empirical distribution functions of the two samples. Its statistic is defined as

$$D_{n,m} = \sup_x \left|F_n(x) - G_m(x)\right|, \tag{15}$$

where F_n and G_m denote the empirical cumulative distribution functions of $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$, respectively. Large values of $D_{n,m}$ indicate increasing discrepancy between the two empirical distributions. Again, because the model parameters are estimated from the same observed sample used for evaluation, standard exact p-values are not formally valid in this setting. Therefore, the significance of the observed KS statistic was assessed using a parametric bootstrap procedure. Combined with the tail-sensitive AD statistic, the two-sample KS test provides a more comprehensive assessment of goodness-of-fit.

The fit quality is further evaluated using the log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta}), \tag{16}$$

where $f(x_i \mid \boldsymbol{\theta})$ denotes the fitted probability density and $\boldsymbol{\theta}$ the parameter vector estimated via maximum likelihood.

Model comparison is performed using the Akaike Information Criterion (AIC) [52] and the Bayesian Information Criterion (BIC) [53], defined respectively as

$$\text{AIC} = -2\,\ell(\boldsymbol{\theta}) + 2k, \tag{17}$$

$$\text{BIC} = -2\,\ell(\boldsymbol{\theta}) + k \log n, \tag{18}$$

where k denotes the number of estimated parameters and n the sample size. Lower AIC or BIC values indicate a more suitable trade-off between model fit and complexity.

Together, the bootstrap-calibrated AD and KS statistics, the log-likelihood, and the information criteria provide a robust quantitative framework for selecting the most representative probability distribution for the observed picking and load/unload time data, ensuring that uncertainty is rigorously incorporated into the scheduling formulation.

5. Experimental results

In this section, the experimental results are presented, analyzing the statistical similarity between distributions across shifts (Section 5.1), the data preprocessing and the performed statistical analysis (Section 5.2), and the Distribution Fitting Analysis and Selection (Section 5.3).

5.1. Shifts analysis

To reveal statistically significant differences in picking-time distributions across shifts for specific components, and to assess whether the shift data are independent and non-aggregable, we applied the Mann–Whitney U test [54] to all pairwise group comparisons. The complete pairwise-comparison results are reported in the Online Supplementary Material (Section S1). Only a small number of pairs failed to reject the null hypothesis of equal distributions at the 5% significance level (p-value > 0.05), indicating limited statistical similarity between those group distributions.

The results suggest that only a few samples exhibit distributions that are statistically similar across shifts. In contrast, data from the same shift across the analyzed months tend to show statistical similarity, indicating temporal consistency in operator performance. These findings support the conclusion that some shift–component combinations can be modeled using a shared distribution. However, due to the significant differences observed across shifts, particularly within the same component, aggregating shift data is not considered an appropriate approach.

5.2. Data preprocessing & statistical analysis

In the absence of reference benchmarks for picking-time durations, and following a detailed analysis of the distributions at the shift level, two modeling scenarios were explored for each component to support subsequent simulation efforts. The first scenario is pessimistic, representing the slowest recorded performance, while the second is optimistic, reflecting the best observed operational performance. These scenarios aim to capture the variability in operator efficiency and provide a realistic range of performance. This dual-scenario modeling ensures that the statistical modeling captures the full spectrum of variability in operator performance and system behavior.

After applying the preprocessing procedure described in Section 4.1 to each shift-level sample, the mean and standard deviation of the cleaned data were calculated, as reported in the Online Supplementary Material (Section S2).

In cases where a shift does not simultaneously exhibit both the highest and lowest mean and standard deviation, the selection is further supported by evaluating the minimum and maximum observed values for each shift. This ensures a more representative identification of performance extremes for use in scenario modeling.

In the pessimistic scenario, the selection is not straightforward. We chose CF_M3 rather than CF_A2 (M stands for May and A for April) because it exhibits the highest mean picking time and a higher minimum, indicating that it is systematically slower across the entire range, not only due to occasional extremes. Although CF_A2 shows a slightly higher maximum and standard deviation, the distribution of CF_M3 is centered further to the right, which is more appropriate for a conservative scenario. Moreover, its larger sample size provides greater confidence that this behavior is stable. The remaining selections met all the defined criteria.

We selected WA_RH_M2, WA_LH_M2, EM_M1, and CF_M1 for the optimistic scenario because they jointly exhibit low mean picking times, relatively small standard deviations, and bounded min–max ranges, indicating fast and consistent performance. Moreover, they belong to the same month (May), ensuring data homogeneity in this scenario. Although CF_A4 shows a lower mean, it belongs to April, and to avoid mixing operational conditions across months, CF_M1 was preferred, given its acceptable dispersion and large sample size.

Table 3
Summary statistics for pessimistic scenario components. WA, LH, RH, EM, and CF denote component identifiers; M = May; 1–4 = shift numbers.

Data	Mean	Std	Min	Max	Length
WA_RH_M3	170.12	51.39	65	327	478
WA_LH_M3	171.57	48.48	68	309	480
EM_M3	167.39	38.10	66	263	457
CF_M3	406.71	66.33	237	567	456

Table 4
Summary statistics for optimistic scenario components. WA, LH, RH, EM, and CF denote component identifiers; M = May; 1–4 = shift numbers.

Data	Mean	Std	Min	Max	Length
WA_RH_M2	124.03	29.25	60	209	516
WA_LH_M2	127.19	29.73	60	210	515
EM_M1	124.77	23.08	60	186	416
CF_M1	335.43	64.94	150	503	511

Table 5
Goodness-of-fit tests and information criteria for the WA_RH_M3 picking-time sample.

Distribution	$\ell(\hat{\theta})$	AD Stat.	AD p	KS Stat.	KS p	BIC	AIC
Gen. normal	−2555	0.2844	0.092	0.0649	0.130	5128	5115
GMM ($N = 2$)	−2542	−0.5379	0.462	0.0502	0.432	5115	5094
GMM ($N = 3$)	−2538	−0.5363	0.334	0.0418	0.560	5126	5092
GMM ($N = 4$)	−2535	−1.0968	0.970	0.0272	0.970	5138	5092
GMM ($N = 5$)	−2535	−0.4644	0.280	0.0439	0.494	5156	5098

5.3. Distribution fitting analysis & selection

After applying the preprocessing procedure described in Section 4.1, the empirical histograms of picking times, together with the corresponding KDE, are reported in the Online Supplementary Material (Section S3) for each selected case, pessimistic and optimistic. Before detailing each case, we consistently report the same set of adjustment metrics: log-likelihood, two-sample AD and two-sample KS test statistics, their associated p-values, and AIC and BIC. The subsequent analysis is performed considering the described pessimistic (Section 5.3.1) and optimistic (Section 5.3.2) scenarios. The final parameterization of the selected models is presented in Section 5.3.3, followed by a literature-based discussion of the fitting results in Section 5.3.4 and implications for synthetic-data generation and scheduling integration in Section 5.3.5.

5.3.1. Pessimistic scenario

Regarding the pessimistic scenario analysis, as summarized in Table 3, we present only the best unimodal distribution that passes the goodness-of-fit tests and minimizes the BIC among the classical statistical distributions considered. This serves as the parsimonious baseline against which the GMMs with $N = 2, \ldots, 5$ are compared. To improve readability, only representative goodness-of-fit tables and graphical diagnostics are retained in the main text, while the complete set of remaining results for the pessimistic scenario is provided in the Online Supplementary Material.

Goodness-of-fit summary for WA_RH_M3. As shown in Table 5, all GMMs yield bootstrap-calibrated two-sample AD and KS p-values above 0.05, indicating an adequate fit. The log-likelihood increases with N, and the AIC is minimal (5092) for $N = 3$ and $N = 4$, whereas the BIC is minimal for $N = 2$ (5115) due to its stronger penalty on model complexity. The KS p-values increase from 0.432 ($N = 2$) to 0.970 ($N = 4$), indicating that $N = 3$–4 provide essentially the best fit, while $N = 2$ remains the most parsimonious.

After inspecting the KDE and Q-Q diagnostics, as illustrated in Figs. 3 and 4, we select the GMM with $N = 3$ components, as it ties for the best AIC (5092), attains bootstrap-calibrated two-sample AD and KS p-values (0.334 and 0.560), yields only negligible log-likelihood gains over $N = 4$, and keeps the BIC lower than more complex mixtures, thereby balancing goodness of fit and parsimony.

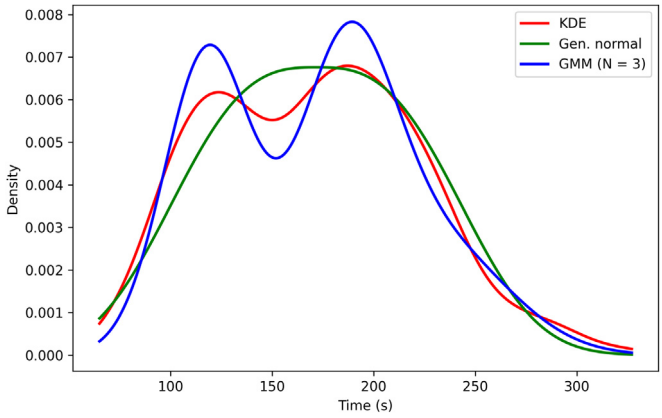

Fig. 3. Empirical density (KDE, red) for WA_RH_M3 overlaid with the best unimodal fit (green, Generalized Normal) and a three-component GMM (blue).

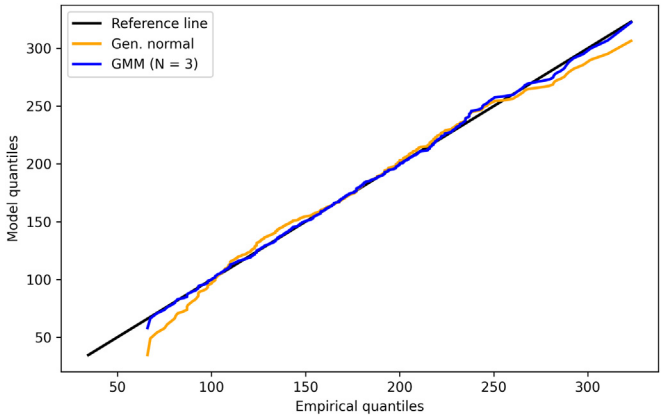

Fig. 4. Q–Q plot for WA_RH_M3 comparing the three-component GMM (blue) and the best unimodal model (orange) against the empirical reference line (black).

Goodness-of-fit summary for WA_LH_M3. All candidate models yield bootstrap-calibrated two-sample AD and KS p-values above $\alpha = 0.05$, indicating an adequate fit. The Beta distribution attains the lowest BIC (5088) and a competitive AIC (5072), together with a high log-likelihood ($\ell(\hat{\theta}) = -2532$), making it the most parsimonious unimodal option. For the GMMs, the log-likelihood increases with the number of components N, and the AIC is smallest for $N = 4$ (5068), although this improvement comes at the cost of greater complexity, reflected in the corresponding BIC increase (5114 for $N = 4$). The KS p-values also improve from 0.316 for $N = 2$ to 0.850 for $N = 3$, indicating that mixtures with $N \geq 3$ provide the closest empirical fit.

After inspecting the KDE and Q–Q diagnostics, we select the GMM with $N = 3$ components as the final model. Although the Beta distribution provides a strong unimodal baseline and closely tracks the overall empirical shape, the three-component GMM more closely reproduces the empirical density across the main mass of the distribution, including the slight shoulder around 120–140 s, while avoiding the over-segmentation introduced by higher-order mixtures. Statistically, the $N = 3$ GMM achieves a high log-likelihood ($\ell(\hat{\theta}) = -2527$), bootstrap-calibrated two-sample AD and KS p-values of 0.398 and 0.850, and an AIC (5070) that remains highly competitive with the minimum AIC obtained for $N = 4$, while retaining a lower BIC (5104 vs. 5114). We therefore consider the $N = 3$ mixture to offer the

best balance between empirical fit and parsimony, and adopt it as the representative picking-time model for WA_LH_M3, while retaining the Beta distribution as a competitive unimodal alternative.

Goodness-of-fit summary for EM_M3. All candidate models except the $N = 3$ GMM yield bootstrap-calibrated two-sample AD and KS p-values above $\alpha = 0.05$, indicating an adequate fit. The log-likelihood increases with the number of mixture components. The AIC is smallest for the $N = 5$ GMM (4619), while the generalized logistic distribution and the $N = 2$ GMM tie at 4622. The BIC is minimal for the generalized logistic model (4634), reflecting its stronger penalty on model complexity. Among the lower-order mixtures, the $N = 2$ GMM attains the strongest combined AD and KS performance, with bootstrap-calibrated p-values of 0.546 and 0.448, respectively.

After inspecting the KDE and Q–Q diagnostics, we select the GMM with $N = 2$ components. Visually, the $N = 2$ mixture closely tracks the empirical density across the main mass and the upper tail, while remaining closer to the reference line in the upper quantiles than the generalized logistic model. Statistically, the $N = 2$ GMM attains bootstrap-calibrated two-sample AD and KS p-values of 0.546 and 0.448, ties the generalized logistic model in AIC (4622), and avoids the additional complexity of higher-order mixtures, whose gains in log-likelihood are not matched by comparable improvements in BIC. We therefore adopt $N = 2$ as the best balance between empirical fit and parsimony.

Goodness-of-fit summary for CF_M3. All candidate models yield bootstrap-calibrated two-sample AD and KS p-values above $\alpha = 0.05$, indicating an adequate fit. The log-likelihood increases with N. The AIC is smallest for $N = 3$, $N = 4$, and $N = 5$ (5117), while the BIC is minimal for the generalized logistic model (5131) and remains lower for the unimodal baseline than for all mixture alternatives. The bootstrap-calibrated AD and KS p-values improve substantially from the generalized logistic model to the higher-order mixtures, with the strongest combined values attained for $N = 4$ and $N = 5$.

After inspecting the KDE and Q–Q diagnostics, we select the GMM with $N = 4$ components. Visually, the $N = 4$ mixture captures the local structure of the empirical density more faithfully than the generalized logistic model, including the mild elevation in the lower range, the main peak, and the secondary rise in the upper tail, while also tracking the reference line more closely in the Q–Q plot. Statistically, it attains bootstrap-calibrated two-sample AD and KS p-values of 0.460 and 0.508, respectively, ties for the best AIC (5117), and improves substantially over the $N = 3$ mixture in both goodness-of-fit measures, while avoiding the additional complexity of the $N = 5$ model. We therefore consider the $N = 4$ mixture to offer the best balance between empirical fit and parsimony, and adopt it as the representative picking-time model for CF_M3.

5.3.2. Optimistic scenario

Likewise, for this scenario, we present only the best unimodal distribution that passes the goodness-of-fit tests and minimizes the BIC among the classical families considered, with ties broken by AIC and then by parameter simplicity. Again, this serves as the parsimonious baseline against which the GMMs with $N = 2, \ldots, 5$ are compared. To improve readability, only representative goodness-of-fit tables and graphical diagnostics are retained in the main text, while the complete set of remaining results for the optimistic scenario is provided in the Online Supplementary Material.

Goodness-of-fit summary for WA_RH_M2. All candidate models yield bootstrap-calibrated two-sample AD and KS p-values above $\alpha = 0.05$, indicating an adequate fit. The bootstrap-calibrated KS p-values improve from 0.150 for the skew-normal model to 0.522 for $N = 2$, while the log-likelihood generally increases with the number of mixture components. Information criteria favor parsimonious models, since the AIC is smallest for $N = 2$ (4922), whereas the BIC is smallest for the skew-normal model (4938). Although the $N = 4$ mixture attains the highest bootstrap-calibrated AD and KS p-values, it does so at a substantially larger BIC (4975), reflecting its higher complexity.

Table 6
Goodness-of-fit tests and information criteria for the WA_LH_M2 picking-time sample.

Distribution	$\ell(\hat{\theta})$	AD Stat.	AD p	KS Stat.	KS p	BIC	AIC
Skew Normal	−2461	−1.0494	0.970	0.0311	0.942	4940	4928
GMM ($N = 2$)	−2459	−0.7512	0.622	0.0388	0.694	4950	4929
GMM ($N = 3$)	−2457	−0.6894	0.460	0.0485	0.308	4964	4930
GMM ($N = 4$)	−2457	−0.6206	0.388	0.0427	0.460	4984	4937
GMM ($N = 5$)	−2455	−0.4133	0.232	0.0563	0.120	4997	4937

After inspecting the KDE and Q–Q diagnostics, we select the GMM with $N = 2$ components. Visually, the $N = 2$ mixture improves upon the skew-normal baseline while remaining essentially comparable to the $N = 3$ mixture in both the density overlay and the Q–Q plot. Statistically, it achieves bootstrap-calibrated two-sample AD and KS p-values of 0.542 and 0.522, respectively, while attaining the best AIC (4922) and avoiding the additional complexity of higher-order mixtures. We therefore adopt $N = 2$ as the best balance between empirical fit and parsimony for WA_RH_M2, while retaining the skew-normal distribution as a strong unimodal baseline.

Goodness-of-fit summary for WA_LH_M2. As shown in Table 6, all candidates yield bootstrap-calibrated two-sample AD and KS p-values above $\alpha = 0.05$, indicating an adequate fit. The log-likelihood increases only mildly with N. The Skew Normal distribution attains both the smallest AIC (4928) and the smallest BIC (4940), while also achieving the highest bootstrap-calibrated AD and KS p-values (0.970 and 0.942, respectively). In contrast, all Gaussian mixtures introduce additional complexity without improving either the information criteria or the goodness-of-fit measures.

After analyzing Figs. 5 and 6 as visual support, we adopt the Skew Normal distribution as the representative model for WA_LH_M2. It closely tracks the empirical density and quantiles while remaining more parsimonious than the mixture alternatives. Since no GMM provides a meaningful improvement in fit relative to the unimodal baseline, the Skew Normal model offers the best balance between empirical agreement, simplicity, and interpretability in this case.

Goodness-of-fit summary for EM_M1. All candidates yield bootstrap-calibrated two-sample AD and KS p-values above $\alpha = 0.05$, indicating an adequate fit. The AIC is smallest for the generalized logistic model, the $N = 2$ GMM, and the $N = 5$ GMM (3797), while the generalized logistic baseline attains the lowest BIC (3809). Among the Gaussian mixtures, the $N = 2$ model provides the strongest combined bootstrap-calibrated AD and KS performance, with p-values of 0.686 and 0.318, respectively, whereas higher-order mixtures offer only limited gains in log-likelihood at substantially larger BIC values.

We adopt the GMM with $N = 2$ components. It ties for the best AIC (3797), improves both the bootstrap-calibrated AD and KS fit relative to the generalized logistic baseline, and remains markedly more parsimonious than the higher-order mixtures. The KDE and Q–Q diagnostics indicate that $N = 2$ tracks the empirical density and quantiles slightly better, particularly in the upper quantiles. Hence, $N = 2$ provides the best balance between empirical fit and parsimony for EM_M1, while retaining the generalized logistic distribution as a strong unimodal baseline.

Goodness-of-fit summary for CF_M1. All candidates yield bootstrap-calibrated two-sample AD and KS p-values above $\alpha = 0.05$, indicating an adequate fit. The generalized normal distribution attains the smallest AIC and BIC (5717 and 5729), but the Gaussian mixtures remain close competitors overall. In particular, the $N = 2$ and $N = 3$ mixtures show goodness-of-fit measures comparable to those of the unimodal baseline, with only small differences across AD, KS, and information criteria.

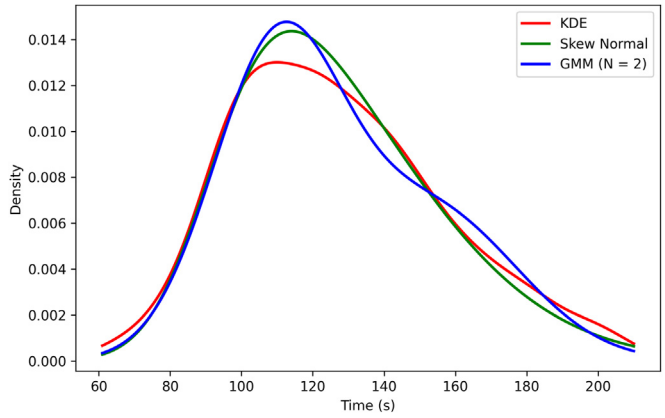

Fig. 5. Empirical density (KDE, red) for WA_LH_M2 overlaid with the best unimodal fit (green, Skew Normal) and a two-component GMM (blue).

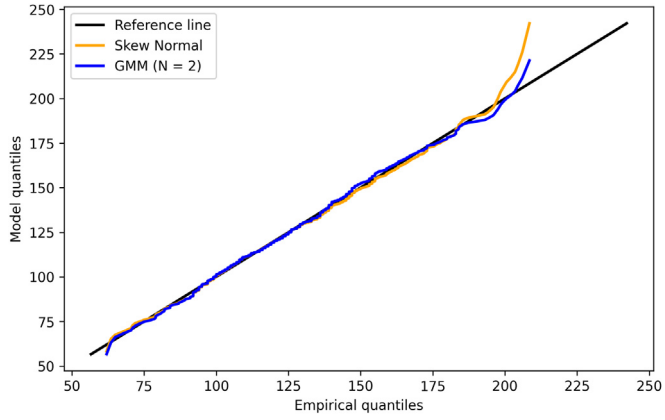

Fig. 6. Q–Q plot for WA_LH_M2 comparing the two-component GMM (blue) and the best unimodal model (orange) against the empirical reference line (black).

We therefore select the GMM with $N = 2$ components as the representative model. Although the generalized normal distribution attains the smallest AIC and BIC, the KDE and Q–Q diagnostics indicate that the two-component mixture captures the mild shoulder in the empirical density slightly better while remaining broadly comparable to the unimodal baseline in overall quantile agreement. Since the information-criterion penalty is modest and the mixture offers a more flexible representation of the observed shape, we consider the $N = 2$ GMM to provide the best balance between empirical fidelity and model complexity in this case, while retaining the generalized normal as a strong parsimonious baseline.

5.3.3. Final parameterization of the selected models

For reproducibility, Table 7 reports the fitted GMM parameters for the cases in which a Gaussian mixture model was selected as the final representative model. For each dataset, components are ordered by increasing mean and shown with mixture weight (π_k), mean (μ_k), variance (σ_k^2), and standard deviation (σ_k). These parameterizations are the ones used when sampling processing and picking times in the DES and when defining scenario-dependent service-time distributions for robust scheduling. The WA_LH_M2 case is not included in this table because its final selected model is the Skew Normal distribution rather than a GMM.

For completeness, the final selected Skew Normal model for WA_LH_M2 was fitted with parameters $(\xi, \omega, \alpha) = (92.355, 45.776, 2.992)$, where ξ is the location, ω the scale, and α the shape parameter.

Table 7
Fitted GMM parameters for the selected picking-time samples, with components ordered by increasing mean.

Dataset	N	Comp.	π_k	μ_k (s)	σ_k^2	σ_k (s)
WA_RH_M3	3	1	0.3855	118.412	461.700	21.49
		2	0.3876	186.084	554.187	23.54
		3	0.2269	230.671	1238.549	35.19
WA_LH_M3	3	1	0.2877	115.568	347.200	18.63
		2	0.4788	176.894	570.907	23.89
		3	0.2335	229.688	1146.956	33.87
EM_M3	2	1	0.6516	150.395	778.038	27.89
		2	0.3484	199.181	1150.293	33.92
CF_M3	4	1	0.1046	290.102	746.233	27.32
		2	0.3000	366.914	597.802	24.45
		3	0.3994	421.320	578.912	24.06
		4	0.1960	500.090	1070.098	32.71
WA_RH_M2	2	1	0.6834	109.655	356.212	18.87
		2	0.3166	155.048	519.760	22.80
EM_M1	2	1	0.6604	114.262	297.772	17.26
		2	0.3396	145.213	353.010	18.79
CF_M1	2	1	0.4450	284.904	2059.203	45.38
		2	0.5550	375.939	2244.782	47.38

5.3.4. Literature-based discussion of the fitting results

The fitting results are consistent with the literature that motivates this study, which highlights that robust intralogistics scheduling depends on a realistic representation of service-time uncertainty rather than on simplified deterministic or unimodal assumptions. In the present case, the results show that picking times cannot be satisfactorily represented across all datasets and operating conditions by a single classical parametric family. Instead, the repeated selection of Gaussian mixture models, particularly in the pessimistic scenario, indicates that the empirical distributions are shaped by multiple latent operational regimes. This finding aligns with the broader literature on industrial intralogistics, where process times are influenced by heterogeneous, context-dependent sources of variability.

A clear pattern is observed across scenarios. In the pessimistic scenario, mixture models were selected in all four cases, with three-component GMMs for WA_RH_M3 and WA_LH_M3, a two-component GMM for EM_M3, and a four-component GMM for CF_M3. In the optimistic scenario, simpler representations were more often sufficient, with two-component GMMs selected for WA_RH_M2, EM_M1, and CF_M1, whereas WA_LH_M2 was adequately represented by a Skew Normal distribution.

The results also show that model selection should not be based solely on information criteria. Although unimodal distributions remained competitive in some cases, especially under optimistic conditions, the selected GMMs generally reproduced the empirical KDE and Q–Q behavior more faithfully, particularly in shoulder regions and tails. This is especially relevant from a scheduling perspective, because inadequate modeling of upper-tail behavior may lead to optimistic duration estimates and, consequently, to less resilient schedules. The preference for low-order mixtures, typically with two or three components, should therefore be interpreted as a compromise between statistical fidelity and parsimony rather than as a purely likelihood-driven choice.

Overall, these findings reinforce the literature-based argument that service-time uncertainty in intralogistics is better represented through flexible, data-driven distributions that capture internal heterogeneity when present, while still allowing simpler models when supported by the data. This provides a statistically and operationally sound basis for the scenario-dependent service-time models later used in the DES and robust scheduling stages of this study.

5.3.5. Integration of the fitted distributions into the scheduling framework

The fitted distributions were not used merely for statistical characterization. Their main role was to define the stochastic parameterization of the service-time components, which were later used in

the DES and scheduling framework. More specifically, the selected scenario-dependent models for picking times were used to represent the variability associated with each component and operating condition, while the load–unload process was represented through the fitted GMM and its empirical approximation. These stochastic service-time terms were then combined with the task-dependent travel-time parameters obtained from the validated simulation model, so that each generated planning instance reflected both the system's physical transport structure and the empirical variability observed in the industrial data.

This integration is central to the scheduling formulation, since the model seeks to minimize waiting times and expected task duration under uncertainty. Accordingly, the fitted distributions directly shape the temporal structure of the optimization inputs rather than remaining as a standalone descriptive analysis. For each task–robot pair, the effective service time depends on the selected picking-time model for the corresponding scenario, together with the sampled load and unload terms and the associated travel-time parameters. In this way, the distribution-fitting stage provides statistically grounded, scenario-sensitive duration inputs for both synthetic data generation and subsequent scheduling experiments.

The practical relevance of this integration becomes clearer when comparing pessimistic and optimistic scenarios. In the pessimistic case, the repeated selection of three-component GMMs indicates more heterogeneous service-time behavior, leading to more dispersed task-duration realizations during planning. By contrast, the optimistic scenario often admits simpler two-component representations, leading to more compact realizations and less dispersed planning instances. Hence, the fitted distributions are operationally relevant because they determine how variability propagates into the DES and MILP stages, ultimately affecting sequencing feasibility, waiting-time propagation, and resource allocation.

A representative example is provided by WA_RH_M3 and WA_LH_M3, for which three-component GMMs were selected, together with EM_M3 and CF_M3, for which two- and four-component GMMs were selected, respectively. When these models are used to generate service-time realizations, the resulting task durations preserve the multimodal or internally heterogeneous structure observed in the empirical data instead of collapsing it into a single average value. The same logic applies to the load–unload process, whose fitted mixture model is sampled when generating planning parameters. Consequently, the scheduling framework is informed by data-driven, scenario-aware service-time inputs, which is precisely the mechanism by which the statistical modeling stage supports the robust scheduling contribution of this study.

Thus, the fitted distributions provide the probabilistic basis for the synthetic service-time generation procedure and for the scenario-dependent parameterization adopted in the scheduling model.

6. Synthetic data generation of service times

This section provides a reproducible procedure for generating synthetic picking and load–unload times for the eight cases comprising the pessimistic and optimistic scenarios analyzed in Section 5. The sampling procedure uses the final fitted models reported in Table 7 and Section 5.3.3 as reference distributions and generates samples in seconds. An optional truncation step excludes values outside the cleaned empirical ranges, thereby ensuring consistency with the preprocessed data.

For a chosen case C among the eight cases presented in Section 5, let the selected model be defined by its final fitted parameters. For the seven cases modeled with GMMs, use the component weights, means, and standard deviations $\{(\pi_k, \mu_k, \sigma_k)\}_{k=1}^{N}$ reported in Table 7. For WA_LH_M2, use the fitted Skew Normal parameters (ξ, ω, α) reported at the end of Section 5.3.3. Let m denote the desired sample size. By default, m is set equal to the corresponding empirical sample size reported in Tables 3 and 4. The optional lower and upper bounds (ℓ, u) may be taken from the corresponding `Min` and `Max` values in these tables in order to constrain the generated samples to the cleaned empirical ranges.

The complete sampling procedure is summarized in Algorithm 1. For the seven GMM-based cases, the sampling step may be fully vectorized by drawing all component labels simultaneously and, for each component k, generating a batch of size $n_k = \#\{i : Z_i = k\}$ from $\mathcal{N}(\mu_k, \sigma_k^2)$, then scattering the results back into $\mathbf{X}$. For WA_LH_M2, the same vectorization principle applies by directly drawing a batch of size m from the fitted Skew Normal distribution.

Algorithm 1 Sampling from the Final Fitted Model with Optional Truncation

Require: Case C, fitted model parameters, sample size m, optional bounds (ℓ, u)
Ensure: Synthetic sample $\mathbf{X}$ in seconds

```
Set pseudo-random seed s
for i = 1 to m do
    if C ≠ WA_LH_M2 then
        Sample Z_i ~ Categorical(π_1, ..., π_N)
        Sample X_i ~ N(μ_{Z_i}, σ²_{Z_i})
    else
        Sample X_i ~ SN(ξ, ω, α)
    end if
    if bounds (ℓ, u) provided then
        while X_i < ℓ or X_i > u do
            if C ≠ WA_LH_M2 then
                Sample Z_i ~ Categorical(π_1, ..., π_N)
                Sample X_i ~ N(μ_{Z_i}, σ²_{Z_i})
            else
                Sample X_i ~ SN(ξ, ω, α)
            end if
        end while
    end if
end for
return X = (X_1, ..., X_m)
```

For the seven GMM-based cases, the optional bounds (ℓ, u) may be replaced, for stress-testing purposes, by mixture-quantile cutoffs. In such cases, the fitted Cumulative Distribution Function (CDF) is given by

$$F(x) = \sum_{k=1}^{N} \pi_k \, \Phi\left(\frac{x - \mu_k}{\sigma_k}\right), \tag{19}$$

where Φ denotes the standard Normal CDF. Choose $\alpha \in (0, 0.5)$, for example $\alpha = 0.01$, and set $\ell = q_\alpha$ and $u = q_{1-\alpha}$, where q_α satisfies $F(q_\alpha) = \alpha$. Because these bounds are induced by the fitted GMM itself, acceptance rates under truncation remain high, and the expected number of resampling operations remains low. For WA_LH_M2, analogous quantile-based bounds may be obtained directly from the fitted Skew Normal CDF using the same definitions of q_α and $q_{1-\alpha}$.

Deterministic MILP models assume a single service time per activity, whereas the empirical data exhibit substantial uncertainty, skewness, and tail behavior. A practical way to incorporate this variability is to replace point estimates with quantiles of the fitted distributions. For each case, we compute the median $q_{0.50}$ and one or more upper quantiles, such as $q_{0.90}$ and $q_{0.95}$. The service time in the MILP can then be set equal to a target quantile according to

$$t^{\mathrm{MILP}} = q_p, \tag{20}$$

so that the probability that the actual service time does not exceed t^{MILP} is approximately p. Alternatively, we can define a tunable buffer around the median according to

$$t^{\mathrm{MILP}}(\lambda) = q_{0.50} + \lambda\big(q_p - q_{0.50}\big), \tag{21}$$

where $0 \le \lambda \le 1$, $\lambda = 0$ recovers the nominal median-based duration and $\lambda = 1$ coincides with the fully robust quantile q_p. The choice of p and λ

modulates robustness, with $q_{0.50}$ reflecting typical conditions and $q_{0.90}$ to $q_{0.95}$ providing protection against upper-tail events at a predictable capacity cost.

For simulation and stress testing, synthetic datasets may be generated using the fitted-model sampling procedure described above. Specifically, draw B independent batches per case using distinct random seeds, run the DES or evaluate the schedule over all batches, and summarize the resulting key performance indicators as empirical distributions, for example by reporting medians together with 5%–95% uncertainty bands.

7. Conclusions & future work

This work develops a statistically principled framework for selecting input-time distributions for AMR picking and load/unload operations. Through outlier-robust preprocessing, nonparametric diagnostics (KDE and Q–Q analysis), and formal goodness-of-fit procedures (two-sample AD and two-sample KS at $\alpha = 0.05$) supplemented with information criteria (AIC and BIC), we determined, for each process and shift, whether classical unimodal families are adequate or whether multi-regime behavior necessitates Gaussian Mixture Models (GMMs). The resulting catalog of distributions and parameterizations is compact, operationally deployable, and faithfully reflects the empirical variability observed in a real production environment.

Although the emphasis of this study lies in statistical inference, the intended application is operational. The selected distributions are designed to interface directly with both DES and MILP models for robust scheduling. In particular, the fitted laws enable (i) expectation- and quantile-based surrogates for deterministic MILP formulations (e.g., using $q_{0.90}$–$q_{0.95}$ as service-time buffers), (ii) chance-constrained or robust counterparts with uncertainty sets calibrated from empirical variance or quantiles, and (iii) scenario generation for stress-testing schedules against tail-event realizations. Collectively, these components yield schedules that are significantly less sensitive to stochastic variability, thereby improving service reliability and resource utilization across the AMR fleet.

An additional benefit is the capacity to support and validate configuration decisions within the plant, such as queue sizing and triggering policies, because the inferred distributions reproduce the observed operational regimes, including the presence of multi-modality, when present system configurations can be re-evaluated under realistic stochastic inputs rather than ad hoc or overly simplified assumptions.

Future research will focus on closing the loop between statistical inference and optimization by (i) embedding the selected distributions within a complete MILP-based scheduling study with explicit service-level guarantees, (ii) extending the present marginal-time models to joint or temporally varying formulations (e.g., conditioned on shift, congestion, or operator effects), and (iii) deploying online re-estimation routines to maintain distributional fidelity over time. Together, these developments will enable the translation of the present distributional insights into robust, high-confidence production schedules.

CRediT authorship contribution statement

Ricardo Moura: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft, Visualization. **Nuno Pessanha Santos:** Methodology, Validation, Writing – Original Draft, Visualization. **Ana Madureira:** Methodology, Software, Formal Analysis, Visualization, Validation. **Alexandra Tenera:** Conceptualization, Validation, Supervision, Project Administration.

Ethics and permissions

Data access and processing followed Volkswagen Autoeuropa's internal data governance policies. Only anonymized operational timestamps were analyzed, and no personal data were accessed or processed.

Author Disclosure of Artificial Intelligence Assistance

The authors acknowledge using an Artificial Intelligence-based (AI-based) language editing tool solely for grammar checking and stylistic proofreading. No part of the conceptual development, analysis, or interpretation of results was generated by AI. All scientific contributions are entirely attributable to the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the support from the AGILE project under the No. LISBOA-01-0247-FEDER-072618. National funds also supported this work through the *Fundação para a Ciência e a Tecnologia* (FCT). Nuno Pessanha Santos received support from the FCT projects LA/P/0083/2020, UIDP/50009/2020, and UIDB/50009/2020 (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020) – Laboratory of Robotics and Engineering Systems (LARSyS). Ricardo Moura received support from the FCT projects UID/297/2025 and UID/PRR/297/2025 – Center for Mathematics and Applications – NOVA Math and Alexandra B. Tenera received support from the FCT project UID/00667/2025 (10.54499/UID/00667/2025).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.dajour.2026.100721.

Data availability

The authors do not have permission to share data.

References

[1] K. Mahmood, K. Karjust, T. Raamets, Production intralogistics automation based on 3D simulation analysis, J. Mach. Eng. 21 (2) (2021) 102–115, http://dx.doi.org/10.36897/jme/137081.

[2] A. Tenera, A. Abreu, A TOC perspective to improve the management of collaborative networks, IFIP, in: Pervasive Collaborative Networks, vol. 283, Springer, 2008, pp. 167–176, http://dx.doi.org/10.1007/978-0-387-84837-2_17.

[3] G. Fragapane, D. Ivanov, M. Peron, F. Sgarbossa, J.O. Strandhagen, Increasing flexibility and productivity in industry 4.0 production networks with autonomous mobile robots and smart intralogistics, Ann. Oper. Res. 308 (1) (2022) 125–143, http://dx.doi.org/10.1007/s10479-020-03526-7.

[4] M. Levorato, D. Sotelo, R. Figueiredo, Y. Frota, Robust permutation flow shop total weighted completion time problem: Solution and application to the oil and gas industry, Comput. Oper. Res. 151 (2023) 106117, http://dx.doi.org/10.1016/j.cor.2022.106117.

[5] D. Lucas, A. Tenera, Input analysis in simulation: A case study based on the variability in manufacturing lines, Adv. Intell. Syst. Comput. 280 (2014) 465–477, http://dx.doi.org/10.1007/978-3-642-55182-6_40.

[6] Y.-J. Yao, Q.-H. Liu, X.-Y. Li, L. Gao, A novel MILP model for job shop scheduling problem with mobile robots, Robot. Comput.-Integr. Manuf. 81 (2023) 102506, http://dx.doi.org/10.1016/j.rcim.2022.102506.

[7] I. Cohen, K. Postek, S. Shtern, An adaptive robust optimization model for parallel machine scheduling, European J. Oper. Res. 306 (1) (2023) 83–104, http://dx.doi.org/10.1016/j.ejor.2022.07.018.

[8] B. Vaisi, A review of optimization models and applications in robotic manufacturing systems: Industry 4.0 and beyond, Decis. Anal. J. 2 (2022) 100031, http://dx.doi.org/10.1016/j.dajour.2022.100031.

[9] T. Afonso, A.C. Alves, P. Carneiro, A. Vieira, Simulation pulled by the need to reduce wastes and human effort in an intralogistics project, Int. J. Ind. Eng. Manag. 12 (4) (2021) 274–285, http://dx.doi.org/10.24867/IJIEM-2021-4-294.

[10] P. Staczek, J. Pizon, W. Danilczuk, A. Gola, A digital twin approach for the improvement of an autonomous mobile robots operating environment: A case study, Sensors 21 (23) (2021) 7830, http://dx.doi.org/10.3390/s21237830.

[11] A.M. Law, Simulation Modeling and Analysis, fifth ed., McGraw-Hill Education, New York, 2013.

[12] M. Schlecht, S. Himmiche, V. Goepp, R. De Guio, J. Köbler, Data-driven decision process for robust scheduling of remanufacturing systems, IFAC-PapersOnLine 55 (10) (2022) 755–760, http://dx.doi.org/10.1016/j.ifacol.2022.09.500.

[13] Imeguisa Portugal, Agile project – resilience, optimization, and efficiency in real-time supply of assembly lines by a fleet of autonomous mobile robots, Technical Annex (2021).

[14] M. Almeida, Estudo do Abastecimento de uma Linha de Montagem: Projeto AGILE (Master's dissertation), Universidade Nova de Lisboa, 2022.

[15] A.R.M. Madureira, Calendarização Robusta de Veículos Autónomos através de Programação Linear Inteira Mista: Aplicação Industrial ao Projeto AGiLE NOVA, (Master's thesis), School of Science and Technology, Universidade Nova de Lisboa, 2024.

[16] E.A. Oyekanlu, et al., A review of recent advances in automated guided vehicle technologies: Integration challenges and research areas for 5G-based smart manufacturing applications, IEEE Access 8 (2020) 202312–202353, http://dx.doi.org/10.1109/ACCESS.2020.3035729.

[17] Y. Lu, The current status and developing trends of industry 4.0: A review, Inf. Syst. Front. 27 (1) (2025) 215–234.

[18] L. Liu, T. Qu, L. Ma M. Thürer, Z. Zhang, M. Yuan, A new knowledge-guided multi-objective optimisation for the multi-AGV dispatching problem in dynamic production environments, Int. J. Prod. Res. 61 (17) (2023) 6030–6051, http://dx.doi.org/10.1080/00207543.2022.2122619.

[19] J. Gao, X. Zheng, F. Gao, X. Tong, Q. Han, Heterogeneous multitype fleet green vehicle path planning of automated guided vehicle with time windows in flexible manufacturing system, Machines 10 (3) (2022) 197, http://dx.doi.org/10.3390/machines10030197.

[20] N. Yasue E. Ruiz Zúñiga, T. Hirose, H. Nomoto, T. Sawaragi, An integrated discrete-event simulation with functional resonance analysis and work domain analysis methods for industry 4.0 implementation, Decis. Anal. J. 9 (2023) 100323, http://dx.doi.org/10.1016/j.dajour.2023.100323.

[21] T. Kopp, M. Baumgartner, M. Seeger, S. Kinkel, Perspectives of managers and workers on the implementation of automated-guided vehicles (AGVs): A quantitative survey, Int. J. Adv. Manuf. Technol. 126 (11) (2023) 5259–5275.

[22] S. Himmiche, A. Aubry, J.-F. Pétin. A framework for robust scheduling under stochastic perturbations P. Marangé, IFAC-PapersOnLine 54 (1) (2021) 1168–1173, http://dx.doi.org/10.1016/j.ifacol.2021.08.137.

[23] Y.-Z. Li, Q.-K. Pan, K.-Z. Gao, M.F. Tasgetiren, B. Zhang, J.-Q. Li. A green scheduling algorithm for the distributed flowshop problem, Appl. Soft Comput. 109 (2021) 107526, http://dx.doi.org/10.1016/j.asoc.2021.107526.

[24] A. Tenera, V.A. Cruz Machado, Critical chain project management: A new approach for time buffer sizing, in: IIE Annual Conference, 2007, pp. 475–480.

[25] S. Aramesh, S.M. Mousavi, V. Mohagheghi, E.K. Zavadskas, J. Antucheviciene, A soft computing approach based on critical chain for project planning and control in real-world applications with interval data, Appl. Soft Comput. 98 (2021) 106915, http://dx.doi.org/10.1016/j.asoc.2020.106915.

[26] W. Peng, X. Lin, H. Li, Critical chain based proactive-reactive scheduling for resource-constrained project scheduling under uncertainty, Expert Syst. Appl. 214 (2023) 119188, http://dx.doi.org/10.1016/j.eswa.2022.119188.

[27] R.L.C. Souza, A. Ghasemi, A. Saif, A. Gharaei, Robust job-shop scheduling under deterministic and stochastic unavailability constraints due to preventive and corrective maintenance, Comput. Ind. Eng. 168 (2022) 108130, http://dx.doi.org/10.1016/j.cie.2022.108130.

[28] Y. Sui, Z. Wang, Parallel identical machines scheduling to minimize the maximum inter-completion time with uncertain processing time, IFAC-PapersOnLine 55 (10) (2022) 2599–2604, http://dx.doi.org/10.1016/j.ifacol.2022.10.101.

[29] M. Panzer, B. Bender, N. Gronau, Neural agent-based production planning and control: An architectural review, J. Manuf. Syst. 65 (2022) 743–766, http://dx.doi.org/10.1016/j.jmsy.2022.10.019.

[30] A. Novak, P. Sucha, M. Novotny, R. Stec, Z. Hanzalek, Scheduling jobs with normally distributed processing times on parallel machines, European J. Oper. Res. 297 (2) (2022) 422–441, http://dx.doi.org/10.1016/j.ejor.2021.05.011.

[31] J. Duan, J. Wang, Robust scheduling for flexible machining job shop subject to machine breakdowns and new job arrivals considering system reusability and task recurrence, Expert Syst. Appl. 203 (2022) 117489, http://dx.doi.org/10.1016/j.eswa.2022.117489.

[32] E. Zimmermann, T.T. Mezgebe, H. Bril El Haouzi, P. Thomas, R. Pannequin, M. Noyel, Multicriteria decision-making method for scheduling problem based on smart batches and their quality prediction capability, Comput. Ind. 133 (2021) 103549, http://dx.doi.org/10.1016/j.compind.2021.103549.

[33] C.G. Corlu, A. Akcay, W. Xie, Stochastic simulation under input uncertainty: A review, Oper. Res. Perspect. 7 (2020) 100162, http://dx.doi.org/10.1016/j.orp.2020.100162.

[34] M. Schleyer, K. Gue, Throughput time distribution analysis for a one-block warehouse, Transp. Res. Part E: Logist. Transp. Rev. 48 (3) (2012) 652–666, http://dx.doi.org/10.1016/j.tre.2011.10.010.

[35] D. Loske, M. Klumpp, E.H. Grosse, T. Modica, Storage systems' impact on order picking time: An empirical economic analysis of flow-rack storage systems, Int. J. Prod. Econ. 261 (2023) 108887, http://dx.doi.org/10.1016/j.ijpe.2023.108887.

[36] M. Vazquez-Noguerol, J.C. Prado-Prado, Application of analytics in food retailing to improve online order picking time estimations, Int. J. Prod. Econ. 280 (2025) 109497, http://dx.doi.org/10.1016/j.ijpe.2024.109497.

[37] E. Guerrazzi, V. Mininno, D. Aloini, Balancing picking and outbound loading efficiency in an SBS/RS through a digital twin, Flex. Serv. Manuf. J. (2024) http://dx.doi.org/10.1007/s10696-024-09554-w.

[38] C. Schumacher, J. Stilling, J. Kriege, P. Buchholz, Live fitting of process data within digital twins of manufacturing to use simulation and optimisation, J. Simul. 18 (5) (2024) 813–834, http://dx.doi.org/10.1080/17477778.2024.2376711.

[39] T. Cokelaer, et al., Cokelaer/Fitter: v1.7.1, Zenodo, 2024, http://dx.doi.org/10.5281/zenodo.12514960.

[40] J.W. Tukey, Exploratory Data Analysis, Springer, 1977.

[41] G. Brys, M. Hubert, A. Struyf. A robust measure of skewness, J. Comput. Graph. Statist. 13 (4) (2004) 996–1017.

[42] M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, Comput. Statist. Data Anal. 52 (12) (2008) 5186–5201.

[43] S. Nadarajah. A generalized normal distribution, J. Appl. Stat. 32 (7) (2005) 685–694, http://dx.doi.org/10.1080/02664760500079464.

[44] A. Azzalini, A class of distributions which includes the normal ones, Scand. J. Stat. 12 (2) (1985) 171–178.

[45] N. Balakrishnan, M.Y. Leung, Order statistics from the type I generalized logistic distribution, Commun. Stat. – Simul. Comput. 17 (1) (1988) 25–50, http://dx.doi.org/10.1080/03610918808812648.

[46] C. Forbes, M. Evans, N. Hastings, B. Peacock, Statistical Distributions, John Wiley & Sons, 2011.

[47] P. Virtanen, et al., SciPy 1.0: Fundamental algorithms for scientific computing in python, Nature Methods 17 (2020) 261–272, http://dx.doi.org/10.1038/s41592-019-0686-2.

[48] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, New York, 2009, http://dx.doi.org/10.1007/978-0-387-84858-7.

[49] P.H. Kvam, B. Vidakovic, Nonparametric Statistics with Applications To Science and Engineering, Wiley-Interscience, 2007.

[50] F.W. Scholz, M.A. Stephens. K sample Anderson-Darling tests, J. Amer. Statist. Assoc. 82 (399) (1987) 918–924, http://dx.doi.org/10.1080/01621459.1987.10478517.

[51] J. Kiefer, K-sample analogues of the Kolmogorov–Smirnov and cramér-von mises tests, Ann. Math. Stat. (1959) 420–447.

[52] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.

[53] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (2) (1978) 461–464.

[54] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. (1947) 50–60.