Big data analytics for quality variation over work shifts in manufacturing systems

Le Toan Duong, Audine Subias, Louise Travé-Massuyès, Christophe Merle

Big data analytics for quality variation over work shifts in manufacturing systems

Le Toan Duong[a,b,c], Audine Subias[a,b], Louise Travé-Massuyès[a,b] and Christophe Merle[b,c]

[a]LAAS-CNRS, Université de Toulouse, Toulouse, France;
[b]ANITI, Université de Toulouse, Toulouse, France;
[c]Vitesco Technologies France SAS, Toulouse, France

ARTICLE HISTORY
Compiled May 17, 2024

ABSTRACT
In the manufacturing industry, mass production enables manufacturers to produce parts with high precision and lower costs. Multiple shifts operate production processes to maximize efficiency. Several researches have been conducted on the impact of shift work on labor's health and habits. However, there have been few studies on the influence of shift work on the consistency of product quality. This paper provides a methodology to analyze the impact of different work shifts in a real electronic board manufacturing industry. The study uses big data and analytics to assess product quality from data. Non-parametric kernel density estimation is used to approximate the distribution of good products in each work slot. Then several metrics are used to measure the dissimilarity between the estimated densities. The approach can be leveraged in various problems related to process performance and quality. The obtained results show that there is no significant difference in terms of product quality between work shifts. These prove the consistency of the manufacturing processes and the homogeneity of performance across work shifts in the studied factory. In a situation in which the results would show a difference, the proposed approach provides valuable information for the company to improve the organization of shift work. Compared to the literature, the paper presents the first quantitative analysis to compare production process performance and product quality over shift works.

KEYWORDS
Big data and analytics; Kernel density estimation; Industry 4.0; Shift work; Production quality

1. Introduction

The manufacturing industry today is a highly competitive market. It becomes even more blatant when the COVID-19 pandemic hits, causing a crisis in the global supply chain and high energy prices (Panwar, Pinkse, and De Marchi 2022). To survive, manufacturers must stay ahead of their competitors. The solution is to offer high-quality and customized products with low cost and short production time (Cadavid et al. 2020). Companies must invest in new technologies to achieve these objectives. At the Hannover Fair 2011 in Germany, the term "Industry 4.0" emerged to describe how to

CONTACT Le Toan Duong. Email: ltduong@laas.fr

create a new industrial revolution in which virtual and physical manufacturing systems flexibly cooperate (Schwab 2017). Among several technologies for Industry 4.0, such as the Internet of Things (IoT), cloud computing, cyber-physical systems (CPS) or artificial intelligence (AI), big data and analytics (BDA) play a crucial role in applying the analysis of the manufacturing big data to extract knowledge and intelligence (Qi and Tao 2018). With the advancement of digital transformation in manufacturing, data are being collected in real-time and automatically at every production stage, including machines, devices, and operators. The comprehensive collection and evaluation of data from different sources support real-time decision-making. Indeed, BDA identifies problems such as bottlenecks or underperforming machines. These problems affect the productivity of the whole process and need to be detected as soon as possible. By analyzing historical data, more accurate predictions can be made to identify possible failures before a breakdown occurs. These are just a few applications of BDA. There are many others in supply chain management (Wang et al. 2016), quality control (Stojanovic et al. 2016), production process optimization (Ungermann et al. 2019; Chou et al. 2005; Lin, Yu, and Chen 2019), etc.

This paper presents a use case of BDA for production process performance analysis at Vitesco Technologies, an automotive manufacturing company that assembles electronic boards for electrified vehicles. Production data are collected in real-time by manufacturing execution systems (MES). These data feed a digital twin, a virtual representation of products and processes. Data processing and analysis allow us to measure product quality and detect abnormal behavior. At Vitesco Technologies, assembly processes run continuously, nearly seven days a week and 24 hours a day. Many of these processes are automatic, and others are performed with the help of humans. In order to satisfy customers, it is essential to ensure that the quality of final products is uniform. However, it may differ in reality due to variations in the quality of the raw materials and the environmental conditions of the factory. It also depends on the performance of different lines and different work teams.

In this use case, the overall performance of the assembly process is analyzed and compared across different time slots during one year. The objective of this study is to verify if there is a difference in performance between the time slots and then to identify the factor that drives these differences. The proposed approach uses a 2D Kernel Density Estimation (KDE) method to approximate the distribution of nominal products for each time slot and then compute their dissimilarity. The results are beneficial for the factory to evaluate the impact of the work shift and work time on production performance. From this point, the company can propose solutions to help improve the quality of work Laosirihongthong, Teh, and Adebanjo (2013).

The paper is organized as follows. A literature review of big data analysis in the context of Industry 4.0 is presented in Section 2. Section 3 describes the industrial problem and the BDA use case. Section 4 presents the methodology used to answer the problem. Results and discussion are presented in Section 5. In the end, Section 6 concludes the study.

2. Literature review

Data analytics is the use of advanced techniques on data to discover patterns, correlations, trends, etc. It intends to help organizations increase efficiency and improve performance by making better decisions. In the context of Industry 4.0, a huge amount of data are collected every second from devices and can be managed and stored by

cloud computing services. Having this profusion of data, along with the increase in computing power, makes the data analytics area crucial for accurate and timely decision making (Günther et al. 2017; Sagiroglu and Sinanc 2013; Cheng et al. 2018). Several researches have been conducted in this area with various analysis methods and use cases. A review of Big Data analysis in Smart Manufacturing, including the most important research roadmaps in Europe related to this topic, is presented in (Nagorny et al. 2017). This study shows that BDA has several use cases with a huge potential exploitation field in Smart Manufacturing, but the growing amount of data poses a challenge for extracting useful information.

Diverse datasets, including structured, semi-structured, and unstructured data, are collected from different sources and of different sizes. Structured data are stored in a predefined format, usually in a table with connected rows and columns. This type of data can be easily processed. (López-Escobar et al. 2012) presented an analysis of the relationship between spectral vibration measurements and the quality of bearings manufactured in an automotive bearing plant. Regarding unstructured data, images from manufacturing processes have been used for online fault detection (Megahed and Camelio 2012; Caggiano et al. 2019). (Kassner et al. 2015) presented an analytic approach that integrates unstructured and structured data around the product life cycle.

Another literature review on Big Data Analytics for manufacturing processes is investigated in (Belhadi et al. 2019). The study highlights that the most prominent challenges addressed by BDA in manufacturing are Quality and Process Control (Q&PC). These challenges are followed by considerations for energy and environmental efficiency, proactive diagnosis and maintenance, as well as safety and risk analysis. The paper presents a specific use case of BDA, focusing on Q&PC for enhancing product quality.

In today's highly competitive market, product quality holds immense significance in the manufacturing industry. The ability to effectively control and improve product quality allows manufacturers not only to survive but also thrive in the market, leading to long-term success. Extensive research has identified various factors that can impact product quality in the manufacturing processes (Lombard, van Waveren, and Chan 2014). These factors encompass raw materials (Fonteyne et al. 2014; Salim and Johansson 2016), equipment (McKone, Schroeder, and Cua 2001), process design (Foehr et al. 2011), human factors (Sgarbossa et al. 2020; Tuli and Manns 2023), and environmental conditions (Waanders et al. 2020). This paper highlights a factor that has received limited attention in the realm of manufacturing: shift work. Shift work entails a work schedule where employees work in rotational shifts, including evenings, nights, and weekends, in contrast to the conventional daytime schedule. Manufacturing facilities often adopt multiple shifts to ensure uninterrupted production.

Many studies have been conducted on the matter of shift work. However, most of them evaluate the impact of this type of workforce management on workers' health (Lowden et al. 2010; Åkerstedt and Wright 2009; Kecklund and Axelsson 2016; Costa 2010; Boivin and Boudreau 2014). There is scarce research in the literature addressing the impact of work shift on production process performance and product quality (Hanna et al. 2008). This study aims at filling this gap. While (Hanna et al. 2008) addresses how the length and number of work shifts affect productivity, this paper presents a production quality quantitative assessment over a set of work shifts in a real electronic board manufacturing process.

3. Problem description

Much like numerous contemporary industries, Vitesco Technologies finds itself positioned at the core of the fourth industrial revolution. In this transformation, data and connectivity enhance the automation of manufacturing processes and improve performance while reducing costs and boosting quality. The present study leverages Big Data analytics to assess the uniformity of the Printed Circuit Board (PCB) assembly processes at a designated production facility within Vitesco Technologies. The assembly process, illustrated in Figure 1, comprises two primary phases:

- Front End assembly (FE): Electronic components are mounted and soldered onto the PCB.
- Back End assembly (BE): Connectors are added to the electronic boards, and the whole is covered by the housing.

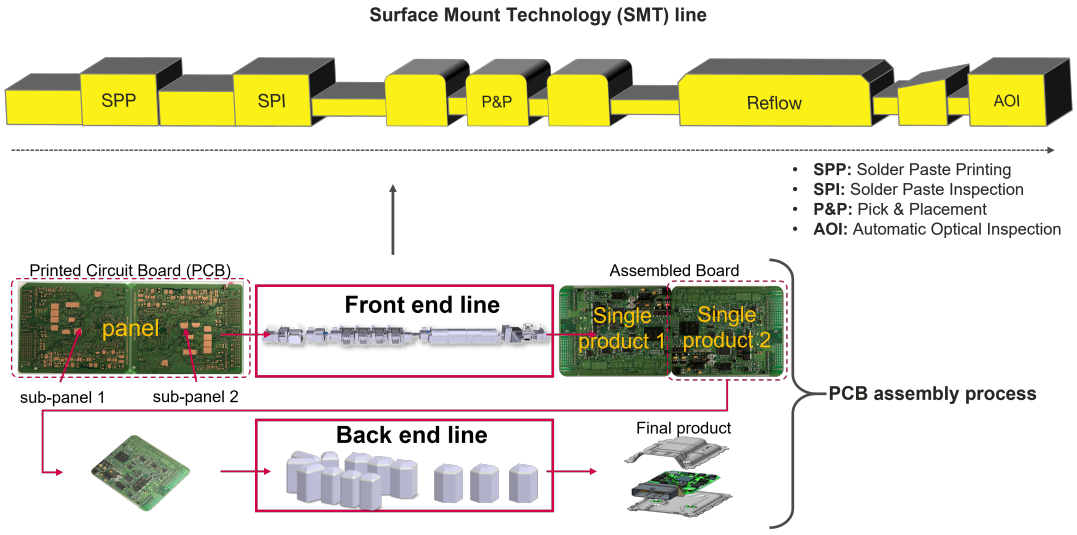


Figure 1.: A schematic view of PCB assembly process

On the Front End assembly, the electronic boards are assembled using surface mount technology (SMT). A set of operations is performed on each side of the PCB. The process for the first side is denoted as $FE1$, and $FE2$ is used for the second side. The Back End phase consists of operations for functional tests and housing. This current study is a continuation of our prior work as presented in (Duong et al. 2021). A more comprehensive understanding of the process can be found within that publication. In our previous work (Duong et al. 2021), a process mining framework was developed to categorize products into different classes based on the deviation of their production path from the nominal path, where a path indicates the sequence of operations and their success or fail status as well as the distribution of time elapsed between operations. This framework employs production data in the form of event logs to construct a process model. Subsequently, the path of products is compared with the nominal process model, which embodies the standard behavior, thereby evaluating their quality across different classes. In that work, the "nominal" class includes all products that follow the nominal path without fail and respect all the time constraints between operations. Having a high proportion of nominal products means that the production process works properly, with little interruption, and the quality of the product is guar-

anteed. Conversely, a low proportion reveals that an important number of products fail at some inspection and test operation or fail to respect time constraints. In such cases, the critical task involves pinpointing the underlying factors contributing to these deviations. Potential factors encompass malfunctioning processes, defects within batches of raw materials, or even a change of work shift. This work focuses on the latter, primarily attributed to the limited extent of research on this aspect within the existing literature. An analysis of production performance through different work time slots is performed to explore this matter. The amount and proportion of "nominal" products produced every 30 minutes are used as performance metrics. A dissimilarity matrix is built using the KDE to compare the performance between time slots. As considered on the production site, the analysis also makes a distinction between weekdays and weekends.

Focusing on the analysis between work slots of $\Delta t = 30$ minutes, the study compares the performance between the 48 production slots of one day, denoted as w_k, i.e., $k \in [1..48]$. For each slot and each day along one year, data are collected and computed to obtain the proportion and number of nominal products. Since the type and number of products produced per day vary according to customer demand, the analysis indeed considers both the proportion and the number of nominal products. Given a work slot w_k, let us define $x_k = (x_k^1, x_k^2) \in \mathbb{R}^2$, a 2-dimensional vector, where x_k^1 and x_k^2 are the variables representing the proportion and the number of nominal products in the work slot w_k, respectively. Daily samples are indexed by j, i.e., $w_{k,j} = (x_{k,j}^1, x_{k,j}^2)$, where $j = 1, \ldots, N_k$, and the set of samples for every slot w_k is denoted by $W_k = \{(x_{k,j}^1, x_{k,j}^2)\}_{1 \leq j \leq N_k}$. One important thing to remember is that the size N_k of W_k is not the same for all slots, given that there is no production for particular time periods on some days. These break times are either expected or not. For example, there is no production from midnight to 5 a.m. every Monday in the studied plant, making these periods expected breaks. An example of unexpected break time is a process breakdown or unplanned maintenance.

The difference in the size of the sets W_k, $k = 1, \ldots, 48$, raises a challenge in defining a dissimilarity metric. The most straightforward approach is to extend the size of all sets to the same maximum size. However, this practice results in a loss of information and could lead to an erroneous analysis. Therefore, the solution to overcome this issue is to compare the two-dimensional distribution estimated from these data sets. The present study uses a non-parametric method: the KDE (Parzen 1962). Then, the dissimilarity is defined by the L_1, L_2, and Jensen-Shannon distances (Fuglede and Topsoe 2004) between obtained densities.

4. Methodology

This section presents a method to compute the difference in performance between work slots. Figure 2 gives the workflow of the proposed method. Firstly, raw data collected by the MES are processed and converted to event logs. Then, following (Duong et al. 2021), a process mining framework is used to construct a process model and to categorize products into different quality classes according to the level of deviation from the nominal path. Whereas the process mining framework is applied to one specific product family in (Duong et al. 2021), in this study, it is extended to all product families produced during one year. From the outputs of the process mining framework, only the nominal quality class is extracted for the performance analysis. The following sections present the KDE method to approximate the distribution of nominal products

for each work slot and then compute their dissimilarity.

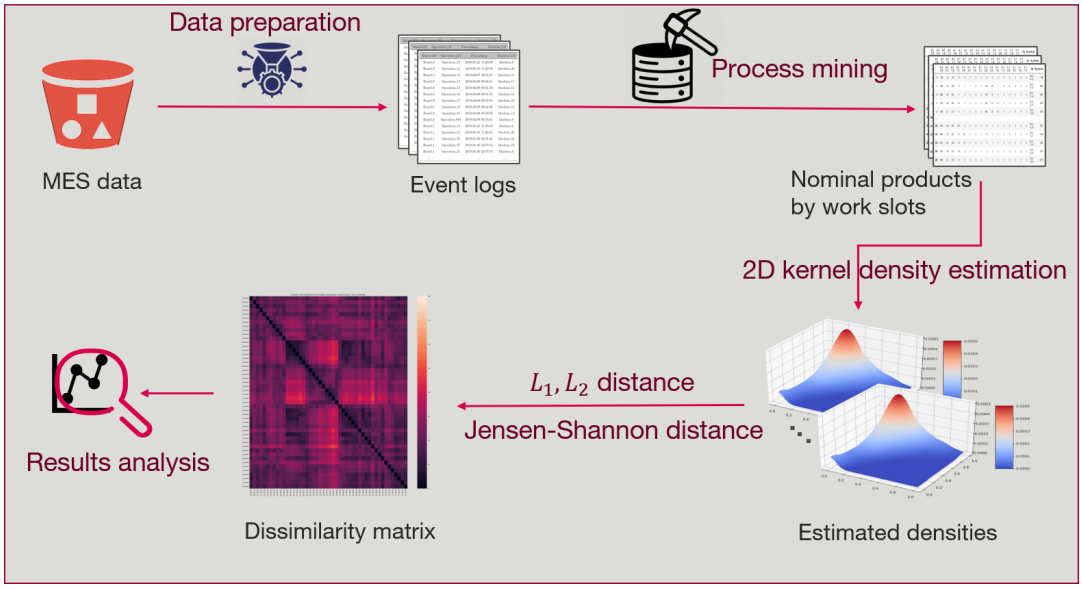


Figure 2.: Performance analysis workflow

4.1. *Kernel density estimation*

In statistics and particularly in statistical estimation, there are two categories: parametric and non-parametric statistics. In parametric statistics, the information about the distribution of a population is known and is associated with a finite set of parameters. Parametric methods are used to estimate these parameters, such as the mean, the variance, etc. On the contrary, non-parametric statistics are either distribution-free or use a specified distribution whose parameters are unspecified.

The KDE method is the most common non-parametric method to estimate the probability density function of continuous random variables. It is also known as the Parzen-Rosenblatt window method (Parzen 1962).

The KDE method was selected due to its flexibility and simplicity. KDE is a non-parametric technique and hence does not require any assumptions about the underlying data distribution. This makes it versatile for modeling complex, multi-modal, or irregularly shaped distributions where parametric methods might fail due to their assumptions. KDE captures the local characteristics of the data distribution, providing a smoother representation that respects the data's variability. The bandwidth parameter, which controls the degree of smoothing applied to the data, enables to control the trade-off between bias and variance, allowing to better tailor the density estimation to the data. The smoothness of the estimated density can help reduce the impact of noisy data. This makes KDE particularly useful when dealing with data that might contain measurement errors or other forms of noise. KDE is relatively straightforward to implement, which makes it a perfect candidate for industrial purposes.

Definition 4.1 (Kernel density estimator of univariate distribution)**.** Consider $\{X_i\}_{1\leq i\leq n}$, a random sample of size n drawn from an unknown probability density

f. The kernel density estimator of f is:

$$\hat{f}_n^h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \qquad \forall x \in \mathbb{R} \tag{1}$$

where K is a kernel and h is the bandwidth or smoothing parameter.

Definition 4.2 (Kernel)**.** A *kernel* is a non-negative real-valued integrable function K such that

- $\int_{-\infty}^{\infty} K(u)du = 1$
- K is an even function, $K(-u) = K(u)$

Some commonly used kernels are given in Table 1.

Kernel	$K(u)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$
Uniform (Tophat)	$\frac{1}{2}\mathbb{1}_{\lvert u\rvert \leq 1}(u)$
Epanechnikov	$\frac{3}{4}(1-u^2)\mathbb{1}_{\lvert u\rvert \leq 1}(u)$
Exponential	$\lambda \exp\left(-\lambda u\right)$
Linear	$(1-\lvert u\rvert)\mathbb{1}_{\lvert u\rvert \leq 1}(u)$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)\mathbb{1}_{\lvert u\rvert \leq 1}(u)$

Table 1.: Common kernels used in KDE

The smoothing parameter h controls the number of samples used to compute the probability for a new point. According to (Parzen 1962) and Turlach (1993), the choice of h is much more important than the choice of K for the behavior of $\hat{f}_n^h(x)$. The quality of the approximation is controlled by the Mean Integrated Squared Error (MISE):

$$MISE(h) = \mathbb{E}_X\left[\|\hat{f}_n^h - f\|_{L_2}^2\right] = \int_{\mathbb{R}} \left(\hat{f}_n^h(x) - f(x)\right)^2 dx \tag{2}$$

$$= \int_{\mathbb{R}} \left(bias^2 \hat{f}_n^h(x) + Var_X(\hat{f}_n^h(x))\right) dx \tag{3}$$

The error is decomposed into two terms. The first term is called bias and the second term is variance. The quality of the estimation depends on the value of the bandwidth parameter h. We have the following properties:

- $Bias \to 0$ when $h \to 0$
- $Variance \to 0$ when $nh \to +\infty$

There is a trade-off between bias and variance. A large window of samples, i.e., a large value of h, may result in a very smooth density with a high bias. In contrast, a small window may have too much detail (high variance) and not be smooth or general enough to correctly cover new or unseen samples. Several approaches were developed to find the optimal value of h. Many of them are based on the assumption that data are sampled from a normal distribution, i.e., Silverman's rule (Silverman 2018), Scott's

rule (Scott 2015), Sheather and Jones method (Sheather and Jones 1991), etc. Cross-validation methods are another technique that does not use any assumptions about the data. These methods aim to fit the model to part of the data and then evaluate the remaining data.

The KDE can be extended to the multivariate case. The most general form is given by Definition 4.3 (Wand and Jones 1994).

Definition 4.3 (Kernel density estimator of multivariate distribution)**.** Consider $\{Z_i\}_{1\leq i\leq n}$, a p-variate random sample of size n drawn from an unknown probability density function $f : \mathbb{R}^p \to \mathbb{R}$. The kernel density estimator of f is:

$$\hat{f}_n^H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H\left(x - Z_i\right) \tag{4}$$

where $x = (x_1, x_2, ..., x_p)^T$ and $Z_i = (Z_i^1, Z_i^2, ..., Z_i^p)^T$. The bandwidth parameter H is a symmetric positive definite $p \times p$ matrix, and

$$K_H(x) = |H|^{-1/2} K(H^{-1/2}x) \tag{5}$$

The matrix H has $\frac{p(p+1)}{2}$ parameters. A simplified version of (4) can be obtained in (6) by choosing the matrix $H = diag(h_1^2, h_2^2, ..., h_p^2)$, allowing different amounts of smoothing in each of the coordinates.

$$\hat{f}_n^H(x_1, ..., x_p) = \frac{1}{n}\frac{1}{\prod_{l=1}^{p} h_l}\sum_{i=1}^{n} K\left(\frac{x_1 - Z_i^1}{h_1}, ..., \frac{x_p - Z_i^p}{h_p}\right) \tag{6}$$

4.2. *Dissimilarity metric for performance analysis*

To measure the dissimilarity between probability distributions, several metrics exist. Among those, both statistical and Euclidean distances are used. The aim is to learn whether the difference in the distances leads to different conclusions. Regarding statistical distance, Jensen-Shannon (JS), a commonly used distance, has been selected. The L_1 and L_2 distances have been used for Euclidean distance. The L_1, L_2 distances between two density probability functions $g(x)$ and $g'(x)$ on a domain S of the euclidean space are defined as:

$$d_{L_1}(g, g') = \|g - g'\|_{L_1} = \int_S |g - g'| dx \tag{7}$$

$$d_{L_2}(g, g') = \|g - g'\|_{L_2} = \sqrt{\int_S (g - g')^2 dx} \tag{8}$$

The JS distance is the root square of the Jensen-Shannon divergence, which is an extension of the Kullback-Leibler (KL) divergence (Fuglede and Topsoe 2004). KL divergence is a very common measure used in probability and statistics that computes a score indicating how much a probability distribution differs from another. Given g and g' the probability density functions of two continuous random variables, their

Kullback-Leibler divergence is given by:

$$D_{KL}(g\|g') = \int_{-\infty}^{+\infty} g(x) \log\left(\frac{g(x)}{g'(x)}\right) dx \tag{9}$$

This measure has two problems. First, it is defined only if $\forall x, g'(x) = 0$ implies $g(x) = 0$. Second, the KL divergence score is not symmetrical, i.e. $D_{KL}(g\|g') \neq D_{KL}(g'\|g)$. The JS distance allows overcoming this later problem because it is symmetrical. The formula of JS distance is as follows:

$$D_{JS}(g\|g') = \sqrt{\frac{D_{KL}(g\|\bar{g}) + D_{KL}(g'\|\bar{g})}{2}} \tag{10}$$

where, $\bar{g} = \frac{g+g'}{2}$.

4.3. *Application to our use case*

This study uses the non-parametric density estimation method KDE because there is no assumption about the data distribution. Formula (6) is applied by choosing H as a diagonal matrix. As defined in section 3, $W_k = \{(x^1_{k,j}, x^2_{k,j})\}_{1 \leq j \leq N_k}$ is the set of data representing the production in a 30-minutes work slot w_k.

According to (6), the 2D probability density function estimated by the KDE for each slot w_k is hence given by $\hat{f}^{H_k}_{kN_k}(x), k = 1..48$, simply denoted by $\hat{f}_k(x)$ for convenience, as follows:

$$\hat{f}_k(x_1, x_2) = \frac{1}{N_k} \frac{1}{h_{k,1} \times h_{k,2}} \sum_{j=1}^{N_k} K\left(\frac{x_1 - x^1_{k,j}}{h_{k,1}}, \frac{x_2 - x^2_{k,j}}{h_{k,2}}\right) \tag{11}$$

where $H_k = diag(h^2_{k,1}, h^2_{k,2})$.

The goal is to assess the dissimilarity of the estimated probability density functions over different time slots.

Computing the dissimilarity of probability density functions solves the problem of sets $(W_k)_{k=1..48}$ having different sizes. Indeed the density $\hat{f}_k(x)_{k=1..48}$ estimated respectively from $(W_k)_{k=1..48}$ can be compared as the dissimilarity between functions.

In this study, the Gaussian kernel is used as the basis function. The Leave One Out (LOO) cross-validation method is used for parameter optimization as the data are not normally distributed and the sample size is small (Sammut and Webb 2011). The probability density function pairs (g, g') used in (7), (8), and (10) for the three dissimilarity metrics L_1, L_2, and the JS distances, respectively, are instantiated with all the possible pairs $(\hat{f}_\kappa, \hat{f}_\nu)$, $\kappa, \nu = 1..48$ and $\kappa > \nu$.

5. Results and discussion

5.1. *Dataset*

In the context of Industry 4.0, data play a central role and are generated at every stage of the production process. Since 2017, Vitesco Technologies has implemented a project to collect data in a structured manner and store them in one location. Figure 3 presents

the cloud architecture developed in this project. The data collection is performed directly from all machines or by the Manufacturing Execution Systems (MES). Then, the NiFi data broker collects and routes data from various sources, such as sensors or log files, to a centralized location. The company uses a cloud storage service such as the one from Amazon to manage these datasets. The collected data, which are stored in Simple Storage Service (S3), are then passed to lambda functions which perform various tasks, such as data transformation and filtering. Finally, the processed data are extracted and executed for further analysis and downstream applications.

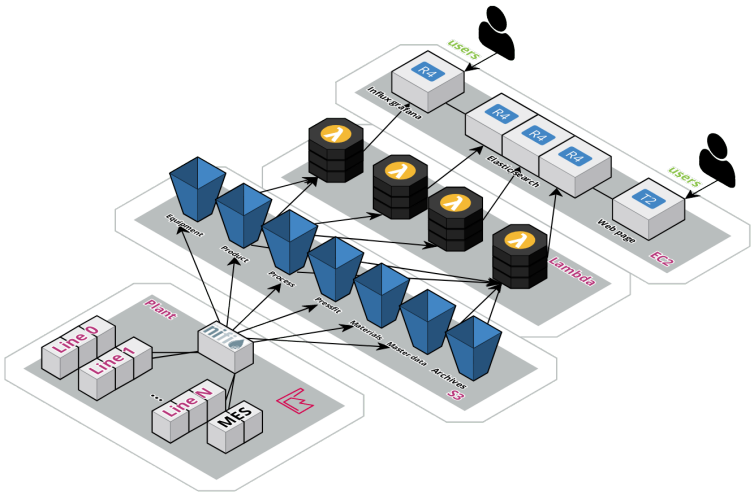


Figure 3.: Cloud architecture for the data collection at Vitesco Technologies.

The data employed in this research pertain to products and encompass event logs detailing the entire production process. All the products at Vitesco Technologies plants are tracked by a unique identifier which is encoded in the form of a data matrix marked on the PCB. This matrix is read each time the PCB goes through an operation. Information from all processes is collected in real-time or near-real-time by the MES and recorded as logs. An example of decoded messages from the cloud storage service of Vitesco Technologies is presented in Figure 4. Each message contains product-related and machine-related information. The main features of the messages are given below:

- **Type of message (red)**: there are several types of messages, among which transitive messages and control messages are mainly used. While transitive messages notify that the PCB has entered or exited some operations, control messages give us information on whether or not the process passed as expected. Hence, these messages have a feature of sanction that could be Pass (P) or Fail (F). An example of a control message is the one generated from the AOI machine.
- **Machine hostname/Machine_ID (pink):** the identification of the machine or computer that performed an operation in the PCB.
- **Description of operation (purple):** description of the performed operation.
- **Family code (green):** product family that is being produced.
- **Serial Number/Board_ID (orange):** the identification number of the product.

- **Operation code/Operation_ID (blue):** the identification number of the operation being performed.
- **Sanction (brown)**: for control messages, the sanction given by the operation (F/P), F means the operation has failed; meanwhile, P letter means the operation has been performed successfully.
- **Timestamp (magenta):** the date and time an operation was performed.

```
P2|mid00002|Ope7_Line2|FamilyC128|Board_560| Operation_7|FACE2||006380|U55H4583|Line2-Config|2|1548866902|
PS|mid00003|Ope4_Line3|FamilyC128|Board_544|1548866935|
SF|mid00000|Ope2_Line0|FamilyC128|Board_566| Operation_2|002|P|1548858285|||||||||||||||||||||||||||||||
SS|mid00000|Ope15_Line0|FamilyC128|Board_730|Operation_15|P|1546860302|
```

Figure 4.: Snapshot of messages file from MES. Data are anonymized to preserve confidentiality.

In a prior study presented in (Duong et al. 2021), these data were utilized to categorize products into distinct classes based on the conformance between their product path with the nominal path. This analysis employed a dataset spanning one year of production activities, comprising over 100 million events associated with approximately 4 million products. For each 30-minute time interval, the number and proportion of nominal products were computed. This study extends the groundwork laid out in the previous work (Duong et al. 2021) by employing various dissimilarity metrics to assess the fluctuations in performance across these time intervals and to investigate the influence of shift work on product quality. This analytical process involves two primary steps. In the initial step, the distribution of nominal products within each time interval is estimated using a set of samples extracted from a year-long timeframe. Subsequently, the dissimilarity matrix for the different time intervals is calculated by quantifying the dissimilarity between the corresponding probability density functions.

5.2. *2D Kernel Density Estimation*

Regarding the density estimation step, the normality assumption on data samples is examined to determine whether parametric estimation methods can be used. For that, the Shapiro–Wilk test (Shapiro and Wilk 1965) is performed on all variables. Figure 5 shows the results of this test. Variables are separated into weekdays (a) and weekends (b), for each phase (*FE1, FE2, BE*), as well as for the quantity represented by the variable (either as the number of nominal products x_k^1 or the proportion x_k^2). Within these tables, each column represents the test result based on the $p - value$ across 48 variables, where $k = 1..48$, corresponding to 48 time slots within a day. If the null hypothesis is rejected, it indicates that the specific variable does not conform to the normal distribution.

This result shows that most variables do not fit the normal distribution, i.e., $p\text{-}value < 0.05$. Then, parametric methods are not applicable in this use case. Consequently, parametric techniques are unsuitable for application in this scenario. As a result, a non-parametric approach is employed. Within the realm of non-parametric methodologies, KDE has emerged as the most suitable for this particular case study, owing to its adaptability and simplicity. KDE offers a versatile and intuitive means of approximating the underlying probability distribution of data, while avoiding the need to make rigid assumptions about its shape. This quality enables it to be well-suited for a diverse array of data types.

	FE1		FE2		BE	
	x_k^1	x_k^2	x_k^1	x_k^2	x_k^1	x_k^2
Null hypothesis rejected (**p-value < 0.05**)	48	41	48	38	48	10
Null hypothesis accepted (**p-value ≥ 0.05**)	0	7	0	10	0	40
Total	48	48	48	48	48	48

(a) Weekday

	FE1		FE2		BE	
	x_k^1	x_k^2	x_k^1	x_k^2	x_k^1	x_k^2
Null hypothesis rejected (**p-value < 0.05**)	48	43	48	40	48	34
Null hypothesis accepted (**p-value ≥ 0.05**)	0	5	0	8	0	14
Total	48	48	48	48	48	48

(b) Weekend

Figure 5.: Results of Shapiro–Wilk test on the normality assumption of data on weekdays (a) and weekends (b)

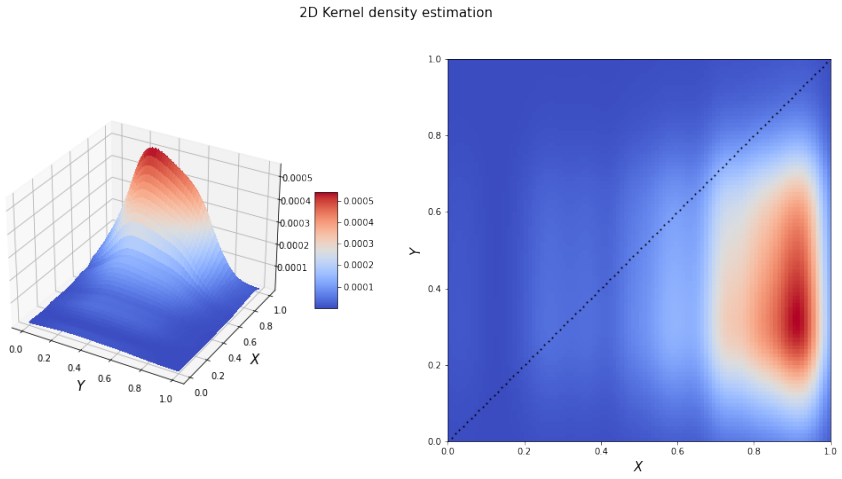
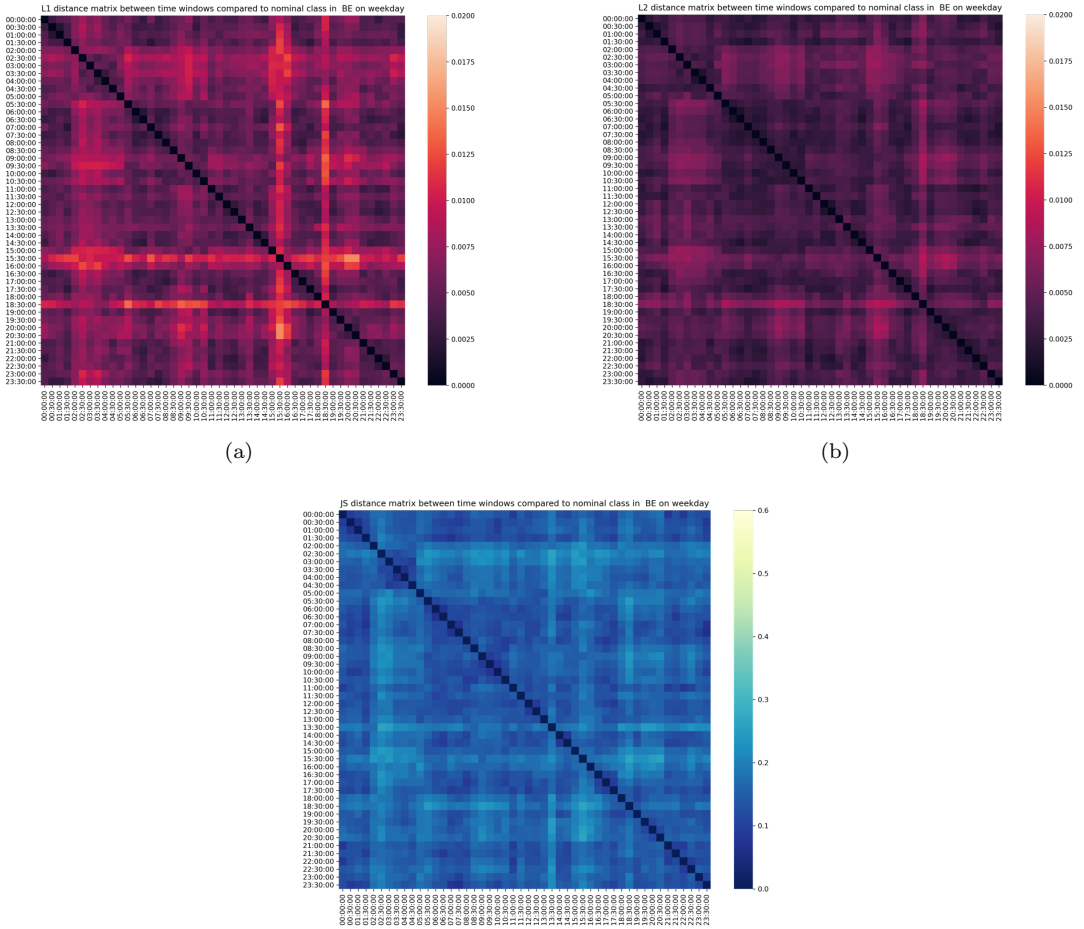


Figure 6.: Estimated density of nominal products in FE1 during time slot [12:00, 12:30] on weekdays. The density is visualized both in 3D plot(**left**) and 2D plot (**right**). The X-axis and the Y-axis represent the proportion of nominal products x_k^1 and the normalized number of nominal products $\tilde{x}_k^2$, with $k = 25$ corresponding to the time slot [12:00, 12:30], respectively.

As mentioned in Section 4, the choice of bandwidth parameter h is much more important than the choice of kernel K. Hence, the study focuses on finding the optimal value of h. For this purpose, the well-known Gaussian kernel is used. As data are not normally distributed, bandwidth selection techniques that rely on a reference distribution are not applicable. In this study, the cross-validation method is used. In particular, the Leave One Out (LOO) cross-validation is used as the sample size is small. The bandwidth parameter h is set among $\{0.03, 0.031, .., 0.1\}$ for the experimentation. The criterion used is the log-likelihood of the data under the estimated model to find the best parameter. Figure 6 presents an estimated 2D density obtained by the proposed method. This is the distribution of nominal products obtained in the $FE1$ phase during the time slot w_k = [12:00, 12:30] on weekdays. The X-axis represents the proportion of nominal products x_k^1, while the Y-axis represents the number of nominal products $\tilde{x}_k^2$ which is normalized by min-max normalization, i.e., $\tilde{x}_k^2 = \frac{x_k^2 - \min\limits_{k=1..48} x_k^2}{\max\limits_{k=1..48} x_k^2 - \min\limits_{k=1..48} x_k^2}$.



(a) (b) (c)

Figure 7.: Dissimilarity matrices between 30 minutes time slots computed from 3 different metrics: L_1(**a**), L_2 (**b**) and JS (**c**).

5.3. *Dissimilarity matrices*

Given a production phase among FE1, FE2, and BE, a work slot w_k is represented by a 2D density $\hat{f}_k$. The dissimilarity matrix $\mathcal{D}$ is a symmetric matrix of size 48×48 where each row or column corresponds to a time slot. Each element $\mathcal{D}_{\kappa\nu}$, $\kappa, \nu = 1..48$, of the matrix is the dissimilarity measure of two corresponding time slots w_κ and w_ν, i.e., $\mathcal{D}_{\kappa\nu} = d(w_\kappa, w_\nu)$.

$\mathcal{D}$ is calculated as follows. First, the exact value of the two estimated densities $\hat{f}_\kappa$ and $\hat{f}_\nu$ are computed at each point of a grid $\mathcal{G}$. The grid is defined in $[0, 1]^2$ with size of 100×100. $\mathcal{D}_{\kappa\nu}$, $\kappa, \nu = 1..48$, are then obtained with the three proposed metrics L_1, L_2, and JS distance, completing the calculation of $\mathcal{D}$. Figure 7 shows the dissimilarity matrices between time slots in the production phase BE on weekdays. The matrix in Figure 7(a) is computed by L_1 metric. Figure 7(b), 7(c) represent the dissimilarity matrices respectively computed by L_2 and JS metric. In all three figures, the darker color represents a lower dissimilarity, and the lighter color shows a higher dissimilarity. As the JS metric does not have the same scale as the L_1 and L_2 metrics, the JS based matrix uses a different color bar to make it more readable.

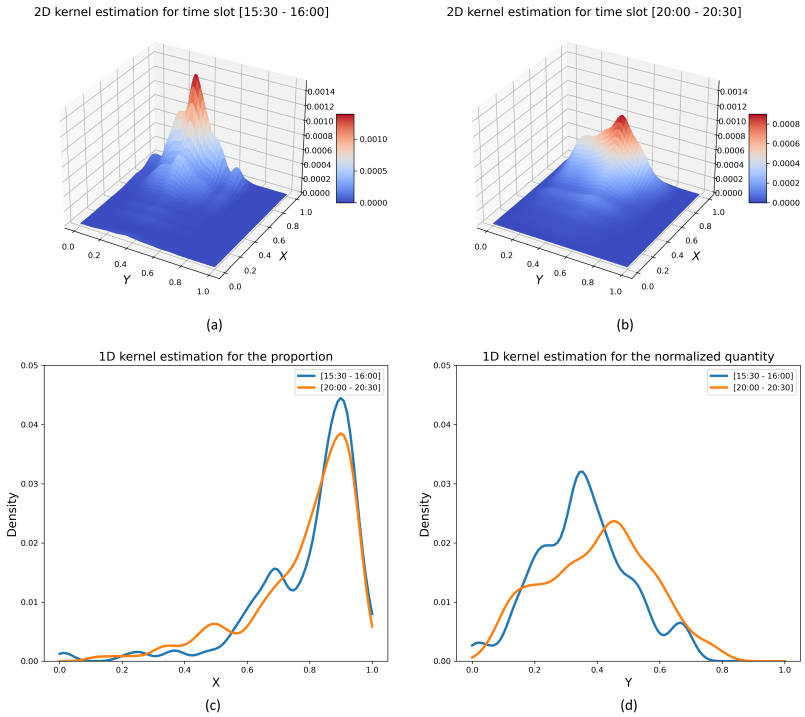


Figure 8.: Visualization of two estimated 2D densities for work slot [15:30-16:00] (**a**) and [20:00-20:30] (**b**) by the KDE method. The X-axis represents the proportion of nominal products. The Y-axis represents the normalized number of nominal products. The 1D densities for each axis X (**c**) and Y (**d**) are also plotted.

An interesting observation highlights that the differences between time slots are more pronounced when employing the L1 distance than the L2 and JS distances. This disparity becomes evident when comparing the dissimilarity matrices of L1 and L2 on the same scale. The L2 distance matrix appears darker due to the involvement of the squaring operation in its calculation. As the discrepancy between two probabilities inherently remains smaller than 1, squaring this value further diminishes its magnitude. Consequently, the choice of metric plays a pivotal role in revealing the dissimilarity between time slots. The L1 distance emerges as the most suitable option in our case study.

Furthermore, the matrix displays several dark blocks, indicative of low dissimilarity. These blocks correspond to neighboring time slots, implying that these proximate slots exhibit more comparable performance in terms of product quality. This is coherent because adjacent time slots are executed under analogous conditions and configurations, potentially linked to the same work shifts. However, the outcomes do not distinctly point out these work shifts. It is essential to remember that on weekdays, there are three work shifts: [5:30, 13:30], [13:30, 21:30], and [21:30, 5:30]. Moreover, the L_1 distance matrix highlights a few slots that significantly differ from the majority of others. These are 15:30, 16:00, and 18:30.

From the engineering point of view, this evidence may lead to managerial decisions.

If the manufactured products are similar over theses slots and the others, then a managerial decision could be to check for existing skill/training gaps in the positions, to review the organization, and to reinforce training. If the manufactured products are different, the manager could think in adjusting the schedules so that the product(s) with the lowest production performance from their team/staff are transferred to the team/staff that has shown better performance, and strengthen training and supervision for the products/teams/staff facing challenges. 15:30, 16:00, and 18:30 are time slots within the same work shift. In this case, if strong production imbalances are shown, then the schedule must be clearly reviewed, and contingency plans (in case of unforeseen events) should be prepared with the teams.

Figure 8 presents the estimated densities of two slots with the largest dissimilarity in the L_1 distance matrix. Figure 8a, 8b present the 2D densities, and Figure 8c, 8d present the 1D densities of each dimension. A key question that arises pertains to determining the degree of dissimilarity that signifies sufficient differentiation between two distributions. The 2D density associated with the slot [20:00-20:30] is flatter, and more spread out than [15:30-16:00]. Regarding 1D density, there is not much difference between the two densities in terms of the proportion of nominal products (X-axis). According to the normalized quantity of nominal products (Y-axis), the distribution that corresponds to the slot [20:00-20:30] is slightly shifted to the right compared to that of [15:00-16:00]. This implies a tendency toward higher values. However, it's important to note that this distinction is not stark. These findings serve as motivation for future work, suggesting the need to develop a metric capable of quantifying the extent of performance differences between different time slots.

6. Conclusion

This paper presents a BDA use case from a real electronic board manufacturing industry. The analysis aims to investigate the quality variability over work shifts. For that, a comparison of distributions of nominal products produced in each 30-minutes slot is carried out. Firstly, the study uses the KDE, a non-parametric method to approximate a probability density from data. Secondly, statistical and general distances are used to compute the dissimilarity between estimated distributions. The findings reveal certain adjacent time slot clusters exhibiting similar performance, although these clusters do not align with the defined work shifts. This suggests that there is no discernible impact of work shifts on product quality. Consequently, these results indicate to the operational team that their focus for performance analysis should shift towards factors other than work shifts. Furthermore, the results identify specific time slots with significantly divergent performance compared to others. This discovery prompts managerial decisions. Overall, the presented work proves the ability of Industry 4.0, in particular, Big Data Analytics, to analyze industrial production processes in support of their optimization. Our proposed approach has a limitation in that it can only identify differences between time slots but cannot determine superiority of one over the other. Therefore, it is not possible to determine which time slot performs better in comparison to others. For future works, a new metric to measure such comparison could be extended. Therefore, in future research, a novel metric could be developed to measure such comparisons. Furthermore, performance analysis could consider additional criteria such as the number of rejected products, failed operations, or average cycle time. Furthermore, the proposed method can be adapted and implemented for other processes in diverse Vitesco Technologies factories.

Acknowledgments

This project is supported by ANITI through the French "Investing for the Future – P3IA" program under the Grant agreement n^{o}ANR-19-P3IA-0004.

Data availability statement

The data that support the findings of this study are not available for confidential issue.

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.

References

Åkerstedt, Torbjörn, and Kenneth P Wright. 2009. "Sleep loss and fatigue in shift work and shift work disorder." *Sleep medicine clinics* 4 (2): 257–271.

Belhadi, Amine, Karim Zkik, Anass Cherrafi, M Yusof Sha'ri, et al. 2019. "Understanding big data analytics for manufacturing processes: insights from literature review and multiple case studies." *Computers & Industrial Engineering* 137: 106099.

Boivin, Diane B, and Philippe Boudreau. 2014. "Impacts of shift work on sleep and circadian rhythms." *Pathologie Biologie* 62 (5): 292–301.

Cadavid, Juan Pablo Usuga, Samir Lamouri, Bernard Grabot, Robert Pellerin, and Arnaud Fortin. 2020. "Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0." *Journal of Intelligent Manufacturing* 1–28.

Caggiano, Alessandra, Jianjing Zhang, Vittorio Alfieri, Fabrizia Caiazzo, Robert Gao, and Roberto Teti. 2019. "Machine learning-based image processing for on-line defect recognition in additive manufacturing." *CIRP annals* 68 (1): 451–454.

Cheng, Ying, Ken Chen, Hemeng Sun, Yongping Zhang, and Fei Tao. 2018. "Data and knowledge mining with big data towards smart production." *Journal of Industrial Information Integration* 9: 1–13.

Chou, Tzu-Chuan, Li-Ling Hsu, Ying-Jung Yeh, and Chin-Tsang Ho. 2005. "Towards a framework of the performance evaluation of SMEs' industry portals." *Industrial management & data systems* .

Costa, Giovanni. 2010. "Shift work and health: current problems and preventive actions." *Safety and health at Work* 1 (2): 112–123.

Duong, Le Toan, Louise Travé-Massuyès, Audine Subias, and Nathalie Barbosa Roa. 2021. "Assessing product quality from the production process logs." *The International Journal of Advanced Manufacturing Technology* 117 (5): 1615–1631.

Foehr, Matthias, Arndt Lüder, Thomas Wagner, Tobias Jäger, and Alexander Fay. 2011. "Development of a method to analyze the impact of manufacturing systems engineering on product quality." In *ETFA2011*, 1–4.

Fonteyne, Margot, Henrika Wickström, Elisabeth Peeters, Jurgen Vercruysse, Henrik Ehlers, Björn-Hendrik Peters, Jean Paul Remon, et al. 2014. "Influence of raw material properties upon critical quality attributes of continuously produced granules and tablets." *European Journal of Pharmaceutics and Biopharmaceutics* 87 (2): 252–263. https://www.sciencedirect.com/science/article/pii/S093964111400071X.

Fuglede, Bent, and Flemming Topsoe. 2004. "Jensen-Shannon divergence and Hilbert space

embedding." In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, 31. IEEE.

Günther, Wendy Arianne, Mohammad H. Rezazade Mehrizi, Marleen Huysman, and Frans Feldberg. 2017. "Debating big data: A literature review on realizing value from big data." *The Journal of Strategic Information Systems* 26 (3): 191–209. https://www.sciencedirect.com/science/article/pii/S0963868717302615.

Hanna, Awad S, Chul-Ki Chang, Kenneth T Sullivan, and Jeffery A Lackney. 2008. "Impact of shift work on labor productivity for labor intensive contractor." *Journal of construction engineering and management* 134 (3): 197–204.

Kassner, Laura, Christoph Gröger, Bernhard Mitschang, and Engelbert Westkämper. 2015. "Product Life Cycle Analytics – Next Generation Data Analytics on Structured and Unstructured Data." *Procedia CIRP* 33: 35–40. 9th CIRP Conference on Intelligent Computation in Manufacturing Engineering - CIRP ICME '14, https://www.sciencedirect.com/science/article/pii/S2212827115006514.

Kecklund, Göran, and John Axelsson. 2016. "Health consequences of shift work and insufficient sleep." *Bmj* 355.

Laosirihongthong, Tritos, Pei-Lee Teh, and Dotun Adebanjo. 2013. "Revisiting quality management and performance." *Industrial Management & Data Systems* .

Lin, Kuo-Ping, Chun-Min Yu, and Kuen-Suan Chen. 2019. "Production data analysis system using novel process capability indices-based circular economy." *Industrial Management & Data Systems* .

Lombard, R., C. C. van Waveren, and K.-Y. Chan. 2014. "Factors affecting quality in a manufacturing environment for a non-repairable product." In *2014 IEEE International Conference on Industrial Engineering and Engineering Management*, 137–142.

López-Escobar, Carlos, Rafael González-Palma, David Almorza, Pedro Mayorga, and María Carmen Carnero. 2012. "Statistical quality control through process self-induced vibration spectrum analysis." *The International Journal of Advanced Manufacturing Technology* 58 (9): 1243–1259.

Lowden, Arne, Claudia Moreno, Ulf Holmbäck, Maria Lennernäs, and Philip Tucker. 2010. "Eating and shift work—effects on habits, metabolism, and performance." *Scandinavian journal of work, environment & health* 150–162.

McKone, Kathleen E, Roger G Schroeder, and Kristy O Cua. 2001. "The impact of total productive maintenance practices on manufacturing performance." *Journal of Operations Management* 19 (1): 39–58. https://www.sciencedirect.com/science/article/pii/S0272696300000309.

Megahed, Fadel M, and Jaime A Camelio. 2012. "Real-time fault detection in manufacturing environments using face recognition techniques." *Journal of Intelligent Manufacturing* 23 (3): 393–408.

Nagorny, Kevin, Pedro Lima-Monteiro, Jose Barata, and Armando Walter Colombo. 2017. "Big data analysis in smart manufacturing: A review." *International Journal of Communications, Network and System Sciences* 10 (3): 31–58.

Panwar, Rajat, Jonatan Pinkse, and Valentina De Marchi. 2022. "The future of global supply chains in a post-COVID-19 world." *California Management Review* 64 (2): 5–23.

Parzen, Emanuel. 1962. "On estimation of a probability density function and mode." *The annals of mathematical statistics* 33 (3): 1065–1076.

Qi, Qinglin, and Fei Tao. 2018. "Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison." *Ieee Access* 6: 3585–3593.

Sagiroglu, Seref, and Duygu Sinanc. 2013. "Big data: A review." In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47.

Salim, Roaa, and Jimmy Johansson. 2016. "The Influence of Raw Material on the Wood Product Manufacturing." *Procedia CIRP* 57: 764–768. Factories of the Future in the digital environment - Proceedings of the 49th CIRP Conference on Manufacturing Systems, https://www.sciencedirect.com/science/article/pii/S2212827116312926.

Sammut, Claude, and Geoffrey I Webb. 2011. *Encyclopedia of machine learning*, 600–601.

Springer Science & Business Media.
Schwab, Klaus. 2017. *The fourth industrial revolution*. Currency.
Scott, David W. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.
Sgarbossa, Fabio, Eric H. Grosse, W. Patrick Neumann, Daria Battini, and Christoph H. Glock. 2020. "Human factors in production and logistics systems of the future." *Annual Reviews in Control* 49: 295–305. https://www.sciencedirect.com/science/article/pii/S1367578820300183.
Shapiro, Samuel Sanford, and Martin B Wilk. 1965. "An analysis of variance test for normality (complete samples)." *Biometrika* 52 (3/4): 591–611.
Sheather, Simon J, and Michael C Jones. 1991. "A reliable data-based bandwidth selection method for kernel density estimation." *Journal of the Royal Statistical Society: Series B (Methodological)* 53 (3): 683–690.
Silverman, Bernard W. 2018. *Density estimation for statistics and data analysis*. Routledge.
Stojanovic, Ljiljana, Marko Dinic, Nenad Stojanovic, and Aleksandar Stojadinovic. 2016. "Big-data-driven anomaly detection in industry (4.0): An approach and a case study." In *2016 IEEE International Conference on Big Data (Big Data)*, 1647–1652.
Tuli, Tadele Belay, and Martin Manns. 2023. "Explainable human activity recognition based on probabilistic spatial partitions for symbiotic workplaces." *International Journal of Computer Integrated Manufacturing* 36 (12): 1783–1800. https://doi.org/10.1080/0951192X.2023.2177742.
Turlach, Berwin A. 1993. "Bandwidth selection in kernel density estimation: A review." In *CORE and Institut de Statistique*, Citeseer.
Ungermann, Florian, Andreas Kuhnle, Nicole Stricker, and Gisela Lanza. 2019. "Data Analytics for Manufacturing Systems – A Data-Driven Approach for Process Optimization." *Procedia CIRP* 81: 369–374. 52nd CIRP Conference on Manufacturing Systems (CMS), Ljubljana, Slovenia, June 12-14, 2019, https://www.sciencedirect.com/science/article/pii/S2212827119303695.
Waanders, Daan, Javad Hazrati Marangalou, Matthäus Kott, Sabrina Gastebois, and Johan Hol. 2020. "Temperature Dependent Friction Modelling: The Influence of Temperature on Product Quality." *Procedia Manufacturing* 47: 535–540. 23rd International Conference on Material Forming, https://www.sciencedirect.com/science/article/pii/S2351978920312154.
Wand, Matt P, and M Chris Jones. 1994. *Kernel smoothing*, 90–92. CRC press.
Wang, Gang, Angappa Gunasekaran, Eric W.T. Ngai, and Thanos Papadopoulos. 2016. "Big data analytics in logistics and supply chain management: Certain investigations for research and applications." *International Journal of Production Economics* 176: 98–110. https://www.sciencedirect.com/science/article/pii/S0925527316300056.