FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Anomaly Detection on Multivariate Time-Series from Lithography Equipment using Machine Learning

João Patrício

Mestrado em Engenharia e Ciência de Dados

Supervisor: Ricardo Sousa (Phd.)

Co-Supervisor: Ana Rita Nogueira (Phd.)

September 27, 2023

Anomaly Detection on Multivariate Time-Series from Lithography Equipment using Machine Learning

João Patrício

Mestrado em Engenharia e Ciência de Dados

Approved in oral examination by the committee:

Chair: Prof. João Moreira (Phd.)
External Examiner: Prof. Carlos Ferreira (Phd.)
Supervisor: Prof. Ricardo Sousa (Phd.)

September 27, 2023

Abstract

The present work is motivated by the challenges embraced by the metallic packing industry in its path along the fourth industrial revolution (Industry 4.0). This work serves as a mean to bring Artificial Intelligence to a mass production lithography process for the detection of anomalous patterns, using Outlier Detection algorithms to support quality control.

Unlabeled time-series from two different lithography production lines (L4 and L8) were collected by Colep at different equipment locations. After a preliminary Exploratory Data Analysis, the datasets were processed through various techniques, from extreme outliers removal to data interpolation.

The processed data was dimensionally reduced using a Principal Components Analysis approach to decrease the problem complexity and allow a visual self-evaluation of the different process operating modes. The production lines L4 and L8 datasets were reduced to two and three principal components, respectively, with a variance explainability above 90%.

As a following step, different clustering models were deployed and evaluated. The best models were achieved using DBSCAN and HDBSCAN for the production lines L4 and L8 datasets, respectively. The different lithography process operating modes were fully mapped, and the Normal Process Behaviour cluster was identified, supported on *Kolmogorov-Smirnov* and non-parametric *Mann-Whitney U* hypothesis tests.

Outlier Detection models were deployed to identify the lithography process outliers and the Normal Process Behaviour was learned. The best results were achieved for both production line datasets when OCSVM-based models were used. For the production line L4 dataset, an F_1 Score and ROC AUC of 99.96% and 99.97% were achieved, respectively. Concerning the production line L8 dataset, an F_1 Score and ROC AUC of 99.83% and 99.80% were achieved, respectively.

Keywords: Industry 4.0, Lithography, Outlier Detection, Time-Series

Resumo

O presente trabalho é motivado pelos desafios enfrentados pela indústria de embalamento metálico, ao longo do seu percurso na quarta revolução industrial (Indústria 4.0). Este trabalho aplica Inteligência Artificial a um processo de litografia, para a deteção de padrões anómalos durante a produção em massa. Deste modo, foram utilizados algoritmos de Deteção de Eventos Anómalos para suportar os controladores de qualidade na tomada de decisão.

Foram recolhidas pela Colep séries temporais não anotadas provenientes de duas linhas de produção (L4 e L8). Após uma Análise Exploratória de Dados preliminar, os conjuntos de dados foram processados utilizando várias técnicas, desde a remoção de eventos anómalos extremos até à interpolação de dados.

Os dados processados foram reduzidos dimensionalmente através de uma Análise de Componentes Principais, de modo a diminuir a complexidade do problema e permitir uma auto-avaliação visual dos diferentes modos de operação do processo. Os conjuntos de dados das linhas de produção L4 e L8 foram reduzidos a dois e três componentes principais, respectivamente, ambos com uma explicabilidade da variância acima de 90%.

Seguidamente, diferentes modelos de agrupamento foram implementados e avaliados. Os melhores modelos foram obtidos por DBSCAN e HDBSCAN para os conjuntos de dados das linhas de produção L4 e L8, respectivamente. Os diferentes modos de operação do processo de litografia foram totalmente mapeados e o agrupamento de dados relativo ao Comportamento Normal do Processo foi identificado, suportado em testes de hipóteses *Kolmogorov-Smirnov* e *Mann-Whitney U*.

Para a identificação de anomalias no processo de litografia, os modelos de Deteção de Eventos Anómalos foram implementados e o Comportamento Normal do Processo foi aprendido. Para ambos os conjuntos de dados da linha de produção, os melhores resultados foram obtidos para modelos baseados em OCSVM. Para o conjunto de dados da linha de produção L4, um F_1 Score e ROC AUC de 99,96% e 99,97% foram alcançados, respectivamente. Em relação ao conjunto de dados da linha de produção L8, foram obtidos um F_1 Score e ROC AUC de 99,83% e 99,80%, respectivamente.

Palavras-Chave: Indústria 4.0, Deteção de Eventos Anómalos, Litografia, Séries Temporais

Acknowledgements

De acordo com a minha contabilização um total de 700 horas foram dispendidas para a realização deste trabalho. Foram, portanto, almoços de domingo que terminaram mais cedo, serões de família que deixaram de existir, viagens de mota que foram canceladas, manhãs de domingo que não foram aproveitadas convenientemente, entre tantas outras vivências do quotidiano que foram colocadas em suspenso.

A vida conforme a conhecia foi colocada em pausa, por um propósito que considero nobre (mas discutível), a aquisição de conhecimento e o aumento das minhas qualificações através da formação. Portanto, antes de efetuar qualquer agradecimento, tenho a obrigação de me desculpar por todas as ausências unilaterais impostas às pessoas que detêm um papel preponderante na minha vida.

Deste modo, primeiramente, gostaria de enaltecer o apoio incondicional dos meus pais pela educação e formação que me foram disponibilizadas, aos dois o meu maior agradecimento! O meu especial obrigado à minha namorada, Sofia, por toda a dedicação, apoio e amor. Não consigo colocar em palavras tudo o que tens feito por mim.

O meu agradecimento aos meus colegas de Mestrado pela partilha de ideias e conhecimento. Um agradecimento muito especial ao Joaquim Carneiro, pelo conhecimento e ideias partilhadas, bem como pela amizade e paciência demonstradas.

Gostaria de agradecer ao meu orientador o Professor Ricardo Sousa e à minha co-orientadora Ana Rita Nogueira por me terem acompanhado neste desafio, pela autonomia e conhecimento transmitido ao longo de todo este projeto.

Por último, esta dissertação foi realizada no âmbito do projeto PRODUTECH4S&C com a referência POCI-01-0247-FEDER-046102.

João Patrício

"We choose to go to the moon!"

John F. Kennedy

Contents

List of Figures

List of Tables

Abbreviations

k-NN k-Nearest Neighbors. 8, 9

AI Artificial Intelligence. 2

APB Anomalous Process Behaviour. 6, 36, 37, 44, 45, 46, 47

CBLOF Cluster-Based Local Outlier Factor. 9

DBSCAN Density-Based Spatial Clustering of Applications with Noise. 10, 11, 32, 33, 55

EDA Exploratory Data Analysis. 17, 19

FPR False Positive Rate. 4, 14

GMM Gaussian Mixture Models. 10, 32, 33

HBOS Histogram-Based Outlier Score. 8

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise. 11, 32, 33, 55

IF Isolation Forest. 8, 44, 45, 46

IQR Interquartile Range. 19

KPI Key Performance Indicator. 3

L4 Line 4. 1, 2, 17, 19, 20, 23, 24, 30, 31, 32, 33, 36, 44, 45, 47, 55

L8 Line 8. 1, 2, 17, 19, 20, 23, 24, 30, 31, 32, 33, 36, 37, 44, 46, 47, 55

LOF Local Outlier Factor. 8, 9, 44, 45, 47

ML Machine Learning. 2, 3, 4, 56

MS Mean Shift. 12, 32, 33

NPB Normal Process Behavior. 4, 5, 7, 36, 37, 44, 45, 46, 47, 55

OCSVM One-Class Support Vector Machine. 9, 44, 45, 46, 47, 56

OD Outlier Detection. 2, 3, 4, 5, 7, 8, 10, 14, 19, 29, 44, 55

OEE Overall Equipment Effectiveness. 3

PCA Principal Component Analysis. 4, 29, 32, 44, 55, 56

ROC Receiver Operating Characteristic. 14, 15

ROC AUC Area Under the Receiver Operating Characteristic Curve. 4, 14, 15, 44, 45, 46, 47, 55

SQL Structured Query Language. 17

TPR True Positive Rate. 14, 15

Chapter 1

Introduction

In this chapter, the work's overall context and application are presented in Section 1.1, followed by the motivations in Section 1.2. The objectives and innovations are proposed in Section 1.3 and the research questions in Section 1.4. In the end, the document structure is described in Section 1.5.

1.1 Context and Application

The proponent Company, Colep, is a metallic packaging production Company based in Vale de Cambra (Aveiro district, Portugal) founded in 1964. The application of the present work is focused on the metallic packaging industry, with particular emphasis on one of its production stages, the lithography process. This critical process is responsible for the final product's aesthetic appearance, as presented in Figure 1.1. The work focuses on two production lines – Line 4 (L4) and Line 8 (L8) – where the lithography process is integrated. A primer varnish application initiates this process over a precut tinplate, where the product's label is printed, and the final product is varnished. Additional details about these three lithography sub-processes are presented below (18).

- **Priming**: a primer varnish and white enamel coating prepare the tinplate's surface for printing to improve the ink's adhesion. The sheet is dried through a gas burner oven, where polymerization occurs together with the fast drying of varnish (18);
- **Printing**: two different printing technologies are available (conventional and ultraviolet), where the product's label is printed (18);
- **Varnishing**: a varnish coating is applied over the printed label as a protective layer. The varnish runs through the rollers until it reaches the tinplate. Various types of varnishes are available, such as primer, white enamel, gold, pigmented, top and matt. After varnishing,

the drying process starts through a gas oven, where the component's polymerization occurs (18).

Currently, all the quality monitoring activities related to the L4 and L8 are manually performed by quality control operators. An example of a non-conforming product is presented in Figure 1.2. The operators are located at the end of each production line, as shown in Figure 1.3, being responsible for visually inspecting each tinplate, assuring that the final product is in accordance with the specifications.

Figure 1.1: An example of a production batch after the lithography process.

Figure 1.2: An example of a non-conforming product due to oven's high temperature.

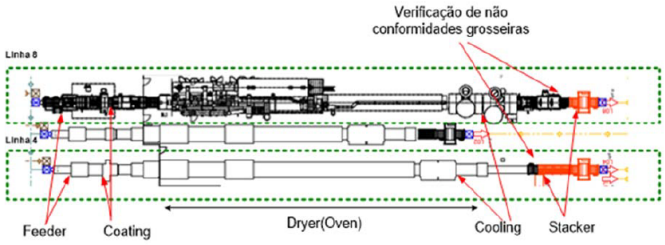

Figure 1.3: L4 and L8 production lines' layouts where the work was developed.

1.2 Motivation

The present work is mainly motivated by the challenges embraced by the metallic packing industry in its path along the fourth industrial revolution (Industry 4.0). This work brings Artificial Intelligence (AI) to a mass production lithography process to detect anomalous patterns, using Outlier Detection (OD) algorithms to support quality control. Some of the OD algorithms deployment are based on Machine Learning (ML) techniques, scratching the surface of the process and quality monitoring applications in industrial scenarios.

Additional environmental and operational motivations are also highlighted in this work (9). The first concerns reducing the production process's ecological footprint by preventing non-

-conformities. The second concerns reducing operational costs, minimizing allocated resources and non-quality costs, and positively impacting the Overall Equipment Effectiveness (OEE) Key Performance Indicator (KPI).

Based on all those mentioned above, we are convinced that the present work will positively impact the brand's trust close to its current business partners and contribute to the Company's competitiveness to attract new ones.

1.3 Objectives and Innovations

As previously introduced, this work's main objective is to build a workflow capable of cataloguing anomalous patterns that potentially correspond to non-conforming production by applying ML techniques. In the end, the overall objective appliance is evaluated by the performance of the OD models deployed based on their evaluation metrics.

Two secondary objectives are also highlighted. The first is to act as proof of concept of a higher complex solution to be deployed (in the future) in this particular metallic packing industry. The second is to act as a decision-making support solution, allowing a prompt root cause analysis and corrective actions prescription by the control operators. However, the nature of these two secondary objectives is primarily intangible. Therefore, an in-depth evaluation of these objectives will not be simple to achieve.

Concerning the innovative aspects of this work and to the best of our knowledge, the proposed approach still needs to be performed in this particular metallic packing industry. Based on this, developing an OD workflow in this mass production process is proof of innovation. Another innovative aspect is the increase of the overall production process insight through the lithography process mapping. Increasing the production process internal knowledge will undoubtedly leverage the shop floor production teams' capabilities and performance.

1.4 Research Questions

To establish a proper scientific approach through the objectives we aim to achieve, two research questions were performed, and a detailed explanation related to the technical approach adopted is described below.

- **How can anomalous lithography production patterns be detected using ML algorithms?**

  Based on our data-specific properties (multivariate time-series), and since no labels or notes from the mass production behaviour were provided during this work, an unsupervised ML approach must be adopted at first.

  Different clustering modelling techniques will be used, and a benchmark will be performed based on internal clustering evaluation metrics. The lithography process will be mapped with the obtained clusters, and the different process operating modes will be identified. This

step is crucial to identify the cluster (or clusters) where the process follows an expected behaviour, defined as the Normal Process Behavior (NPB).

In resume, the clustering modelling and the lithography process mapping will identify (label) the NPB. In the end, different OD algorithms will be used. A benchmark will be performed based on the external clustering evaluation metrics, and the learned frontiers (separation between inliers and outliers) will be presented and discussed.

This research question is crucial since it will be the primary phase for relating the detected outliers with the actual non-conforming batches.

- **What is the OD performance when ML algorithms are applied?**

  A combination of two approaches will be adopted to evaluate the different models' performance. The first uses internal clustering validation since no ground-truth labels are available. Four internal evaluation metrics will be used, namely, the Silhouette index, Davies-Bouldin index, SD index and the S_Dbw index, due to their robustness to monotonicity, noise, density, and skewed distribution scenarios.

  The second uses external clustering validation using the clustering modelling and the lithography process mapping results. Four external evaluation metrics will be used, namely, Recall, False Positive Rate (FPR), F_1 Score and Area Under the Receiver Operating Characteristic Curve (ROC AUC). Although only the latter two were used for the benchmark.

To address the research questions above, *Python* programming language was used, together using well-known libraries such as *Matplotlib*, *NumPy*, *Pandas*, *Scikit-Learn* (16) and *SciPy* (32), together with other customized functions.

1.5 Document Structure

This work is divided into four different Sections:

- **Chapter 2**: provides an introduction to the clustering and OD subjects, describing the different algorithms applied. The related evaluation metrics (internal and external) are also presented;
- **Chapter 3**: describes the data acquisition process and exhaustively details the data preparation process;
- **Chapter 4**: describes all the data modelling methods adopted. Starting with the Principal Component Analysis (PCA), the best clustering models obtained and the lithography process mapping. In the end, the best OD models are presented, evaluated and discussed;
- **Chapter 5**: presents the main results obtained, further work and improvements.

Chapter 2

State of the Art Review

In this chapter, an introduction to the OD subject will be presented in Section 2.1, together with an outliers' taxonomy in time-series in Section 2.2. A list of characteristics for a successful OD model deployment will be detailed in Section 2.3, along with a brief explanation of the relevant algorithms used for this task in Section 2.4. An introduction to the clustering algorithms used to identify the different lithography process operating modes will be described in Section 2.5. At the end of the chapter, the adopted evaluation metrics will be presented in Section 2.6.

2.1 Outlier Detection

In the literature, three significant terms are used to describe nonconforming data patterns: anomalies, outliers and novelties. These terms originated from different application domains without a universally accepted definition (17).

Anomalies and outliers are often used interchangeably in the context of OD to describe data patterns in the feature space that do not follow the expected behaviour (4). In a different domain, novelties are related to previously unobserved patterns in the data (17). For uniformity, the terms outlier and outlier detection (OD) will be adopted in the mathematical context. The term anomaly and anomaly detection will be adopted in the empirical context (e.g., non-conformities are anomalies in the production process).

Although OD approaches are used in various application domains, from credit card fraud detection to cyber-security intrusion detection, we are particularly interested in its industrial applications for process monitoring (33).

An example of how inliers and outliers can be perceived in a process and quality monitoring scenario is presented in Figure 2.1. The example shows two major inlier regions, N_1 and N_2, representing most observations where the process follows an expected behaviour. These regions were also denoted as NPB clusters. On the contrary, points, o_1 and o_2, and region, O_3, are located

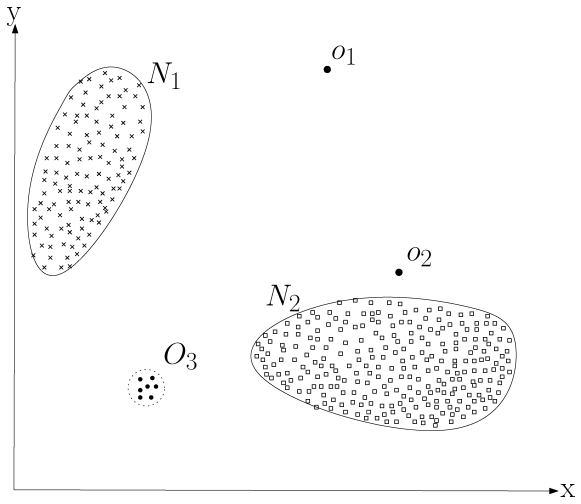

Figure 2.1: An example of inliers and outliers observations in a two-dimensional feature space from Chandola *et al.* (4).

away from inlier regions and therefore considered outliers. These regions were also denoted as Anomalous Process Behaviour (APB) clusters.

From the process monitoring point of view, outliers can be induced by different factors, namely, different operating modes, faulty conditions, process parameters modification, structural changes, and start-up and shutdown periods, to cite some examples (33).

2.2 Outliers in Time-Series

For most process monitoring applications, it is assumed that data instances have a temporal dependency. Therefore, a more specific time-series outlier taxonomy must be explored, where outliers are primarily divided by point- and pattern-wise as described below (10).

- **Point-wise outliers**: when the occurrence is restricted to an individual instance as a glitch or spike.
  - **Global outliers**: individual instances that significantly deviate from their neighbouring values, usually in the form of a spike in the time-series;
  - **Contextual outliers**: individual instances that deviate from their neighbouring context within a time period, usually in small glitches in the time-series.
- **Pattern-wise outliers**: when the occurrence occurs in a subsequence level, in the form of inharmonies in the time-series.
  - **Shapelet outliers**: subsequences with dissimilar basic shapelets compared with the normal shapelet;
  - **Seasonal outliers**: subsequences with unusual seasonalities compared with the overall seasonality;

  - **Trend outliers**: subsequences that significantly alter the time-series trend, resulting in a permanent shift of its mean.

2.3 Outlier Detection Requirements

Since an industrial application for process and quality monitoring is desired, the successful OD candidate(s) should gather a set of essential requirements, which are described below (4; 33).

- **Unsupervised learner**: since the labelling process is expensive, time-consuming and, in some situations, impossible to obtain, most of the data available in industrial scenarios is unlabeled. Moreover, as presented in Section 2.1, outliers can be caused by different factors, making modelling all possible behaviours extremely difficult. On the contrary, inlier data is widely available, being easy and inexpensive to obtain;

- **Real-time operation**: higher output rates might be standard depending on the type of process. Therefore, time-intensive OD models (e.g., distance-based) might face some difficulties. Real-time operation allowance is also intrinsically related to the model's computational complexity;

- **Class imbalance robustness**: since inlier data is widely available compared to outlier data, a natural dataset imbalance will occur. Therefore, the selected model must be capable of reporting a good performance under this scenario;

- **Noise robustness**: because industrial applications have increased in complexity, noisy data similar to outliers might be inherent during data acquisition. Based on this, a deterioration of the model's performance is expected, depending on the degree of sensitivity to noise.

Based on the unsupervised nature of our data and since no labels or annotations were provided during this work, clustering techniques assumed an important role during the modelling phase (Section 4.2). Clustering was used to map (label) the different lithography process operating modes and identify the NPB cluster. This information was then used with the OD algorithms to identify lithography process outliers. A description of the clustering algorithms used is presented in Section 2.5.

2.4 Outlier Detection Algorithms

Based on the specificities of our data, we found mathematical support for the proposed solution on six different unsupervised OD algorithms, which we are going to describe in more detail in the following sub-sections (22; 34).

2.4.1 Isolation Forest

Isolation Forest (IF) (12) is a tree-based ensemble algorithm that takes advantage of two outliers' properties. First, outliers are always the minority class in an OD problem. Second, outliers' attribute values are different compared to the inlier instances. This principle makes outliers easier to isolate, presenting shorter path lengths to the tree root. An association between path length and outlier score can then be performed to identify outliers from the inlier samples.

From the hyperparameters available to fine-tune the model, two were considered in the grid-search, while the others were kept according to their standard values (16; 27):

- **Contamination**: the proportion of outliers in the dataset. The amount of contamination to be used to define the threshold on the samples' scores;
- **Number of estimators**: the number of estimators (tress) that will make part of the isolation forest model.

2.4.2 Histogram-Based Outlier Score

Histogram-Based Outlier Score (HBOS) (7) is a distance-based algorithm that assumes independence between attributes to compute outliers scores based on histograms. Most inliers belong to high-frequency bins, while outliers are related to low-frequency bins. Therefore, the bins' height can be seen as a measure of outlyingness. HBOS can formally be defined as the sum of each attribute's logarithmic univariate outlier score, where higher values will be assigned as outliers.

2.4.3 k-Nearest Neighbors

k-Nearest Neighbors (k-NN) (19) is a distance-based algorithm that computes the distances between each instance in the feature space. For each instance, the pair of distances are calculated and sorted from the lowest to the highest value. First, the lower k distances are selected, where k is a user-defined hyperparameter related to the number of neighbours to be considered. The outlier score of an instance is then defined by the distance to its k nearest neighbour, where instances with a higher outlier score are considered outliers.

2.4.4 Local Outlier Factor

Local Outlier Factor (LOF) (3) is a distance-based algorithm that is particularly efficient when data presents variations in density along the feature space. LOF is the ratio between the density of the neighbourhood where an instance is positioned (instance-oriented) and the density of its closest cluster. Instances considered part of a cluster will have LOF values close to 1. Instances with a LOF value greater than 1 will not be considered part of a cluster and, therefore, will be assigned as outliers.

From the hyperparameters available to fine-tune the model, three were considered in the grid-search, while the others were kept according to their standard values (16; 28):

- **Contamination**: the proportion of outliers in the dataset. The amount of contamination to be used to define the threshold on the samples' scores;
- **Novelty**: defines the approach that will be used by the algorithm (outlier or novelty detection);
- **Number of neighbours**: the number of neighbours to be considered in the *k*-NN calculation.

2.4.5 Cluster-Based Local Outlier Factor

Cluster-Based Local Outlier Factor (CBLOF) (8) is a distance-based algorithm that can be seen as a cluster-oriented version of the LOF algorithm. CBLOF considers the size of the cluster to which an instance belongs, together with the distance between it and its closest cluster. First, instances are assigned to one specific cluster using a clustering algorithm. Then, the obtained clusters are sorted according to their cluster size from a larger to a smaller size, from where the more considerable clusters are selected based on a user-defined hyperparameter threshold. Finally, the distances between instances (from the smaller clusters) to the major clusters' centroids are computed, and an outlier score is assigned. Outliers might not only be considered single instances but also small clusters of instances.

2.4.6 One-Class Support Vector Machine

The One-Class Support Vector Machine (OCSVM) (23) uses a kernel function to map the training data into a higher-dimensional feature space, where the data separation might be easier to achieve. This approach allows the search of a hyperplane that maximizes the margin around the origin of the feature space while containing as many normal data points as possible on the right side of the hyperplane. The data points closest to the hyperplane are considered *support vectors*, having a critical role in defining the decision boundary (learned frontier). During testing, the unseen data will be projected into the same higher-dimensional feature space, and the distance between a new data point and the decision boundary will be calculated. A data point is considered an outlier if it is located outside the decision boundary where the outlier data is expected to be.

From the hyperparameters available to fine-tune the model, one was considered in the grid-search, while the others were kept according to their standard values (16; 30):

- **Nu**: the proportion of outliers in the dataset. The amount of contamination to be used to define the threshold on the samples' scores.

2.5 Clustering Algorithms

To identify the different lithography process operating modes, we found mathematical support on four different clustering algorithms, which we will describe in more detail in the following sub-sections (24).

2.5.1 Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (25) is built on the theory that clusters are high data density areas split by low data density areas. This principle allows DBSCAN to identify clusters with different shapes in addition to the simpler convex-shaped ones. The DBSCAN algorithm is based on the concept of core samples, which are samples located in high data density areas. A cluster is defined by a set of core samples (close to each other) and non-core samples. Non-core samples are not sufficiently close to the core samples to be considered part of the cluster's core.

According to the DBSCAN theory, a high data density area can be described by two hyper-parameters, namely, epsilon and the minimum number of samples. The epsilon controls the local neighbourhood of the points. A lower epsilon value will make most of the dataset noise (labelled as -1). Conversely, higher epsilon values will induce a complete dataset aggregation, resulting in a single cluster model. The minimum number of samples controls the algorithm's sensitivity towards noise.

Based on these two hyperparameters, a core sample is then defined as a sample in the dataset such that there is a minimum number of samples within an epsilon distance, defined as neighbours of the core sample. A cluster can then be identified by iteratively taking a core sample, finding all of its neighbours that are core samples, finding all their neighbours that are core samples and so forth. Ultimately, the non-core sample will be located on the fringes of the cluster.

A sample not considered a core sample located beyond an epsilon in distance from any core sample is considered an outlier. From the hyperparameters available to fine-tune the model, three were considered in the grid-search, while the others were kept according to their standard values (16; 25):

- **Distance**: the metric used for calculating the distance between samples in the feature space;
- **Epsilon**: the maximum distance between two samples for one to be considered as in the neighbourhood of the other;
- **Minimum number of samples**: the number of samples in a neighbourhood for a sample to be considered a core sample.

2.5.2 Gaussian Mixture Models

For real-life applications, the Gaussian model presents one of the simplest approaches to OD. Since its assumptions (unimodal and convex nature of the data) might be easily violated, a linear combination of normal distributions can be adopted to create the Gaussian Mixture Models (GMM) algorithm, according to (2.1),

$$\rho_{MoG}(x) = \frac{1}{N_{MoG}} \sum_j a_j \rho_N \left(x; \mu_j, \sum_j \right) \tag{2.1}$$

with, N_{MoG}, a hyperparameter which defines the number of Gaussians, a_j, the mixing coefficients, ρ_N, the probability distribution, μ_j, the individual mean and, $\sum_j$, the individual covariance (31).

From the hyperparameters available to fine-tune the model, two were considered in the grid-search, while the others were kept according to their standard values (16; 26):

- **Number of components**: the number of mixture components;
- **Type of covariance**: the type of covariance parameters to use.

2.5.3 Hierarchical Density-Based Spatial Clustering of Applications with Noise

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (24) is the hierarquical approach of the DBSCAN algorithm. While DBSCAN assumes that clusters present the same data density (global homogeneity), HDBSCAN mitigates this assumption by exploring all the possible density scenarios.

The algorithm is initiated by calculating the mutual reachability distance between data points, which contains the density of points and their distance from each other. A minimum spanning tree is extracted from the hierarchical representation of the data, which connects the data points based on their mutual reachability distances. HDBSCAN then extracts clusters from the hierarchy by considering different levels of density reachability. Ultimately, a stability score is assigned to each cluster, quantifying the cluster's frequency in different subsamples of the dataset. This approach improves the cluster's robustness, improving the algorithm's sensitivity towards noise. Data points located outside stable clusters are considered noise.

Two hyperparameters are highlighted to achieve optimal performance: the minimum number of samples in the neighbourhood and the minimum cluster size. The former allows HDBSCAN to perform clustering through different data density areas, excluding the need to select an epsilon. The latter defines the minimum number of samples a cluster must have, not to be considered noise.

From the hyperparameters available to fine-tune the model, four were considered in the grid-search, while the others were kept according to their standard values (14; 16):

- **Cluster epsilon**: a distance threshold, where clusters that fall beneath it will be combined;
- **Distance**: the metric used for calculating the distance between samples in the feature space;
- **Minimum cluster size**: the smallest quantity of samples within a group required for that group to be considered a cluster;
- **Minimum number of samples**: the number of samples in a neighbourhood for a sample to be considered a core sample.

2.5.4 Mean Shift

Mean Shift (MS) (24) applies a hill-climbing optimization algorithm, which iteratively shifts a kernel function (flat, Epanechnikov, Gaussian, or other) to a higher-density region of the feature space until convergence is reached. The kernel typically has a bandwidth hyperparameter, which defines the size of the neighbourhood to be considered. In each iteration, the kernel is shifted to the point's mean based on a mean shift vector pointing towards a maximum density region. Depending on the kernel used, different methods are available to compute the points' mean. The algorithm convergence is reached when a shift cannot accommodate more points inside the kernel.

From the hyperparameters available to fine-tune the model, one was considered in the grid-search, while the others were kept according to their standard values (16; 29):

- **Bandwidth**: related to the kernel function, is the size of the neighbourhood to be considered.

2.6 Evaluation Metrics

After fitting the model, inliers and outliers instances are expected to be well separated in the form of clusters in the feature space. For evaluation purposes, two approaches will be explored, depending on the availability of ground truth labels. If no labels are available, an internal cluster validation is performed to evaluate clusters' compactness and separation (13). If ground truth labels are available, an external cluster validation will be first considered (2). Additional details concerning the selected clustering validation metrics are described below.

- **Internal clustering validation**: considered when no ground truth labels are available. Although various validation metrics are available, the selection process considered their robustness to monotonicity, noise, density, and skewed distribution scenarios (13; 6).

  - **Silhouette index**: computed based on the pairwise difference within- and between-cluster distances (21). The within-cluster mean distance, $a(i)$, is defined as the mean distance of a point to the neighbouring points of the cluster it belongs to. The between-cluster distance, $b(i)$, is the smallest mean distance between a point and the points from the other clusters. Thus, for each point, the silhouette width, $s(i)$, is computed according to (2.2).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{2.2}$$

    The cluster mean silhouette, s_k, is then defined as the mean of the silhouette widths of a given k cluster, according to (2.3), where n_k is the total number of points in the k cluster.

$$s_k = \frac{1}{n_k} \sum_{i \in I_k} s(i) \tag{2.3}$$

Finally, the mean of the mean silhouettes through all the clusters is defined by the global silhouette index and is determined according to (2.4), with K the total number of clusters.

$$Global\ silhouette\ index = \frac{1}{K}\sum_{k=1}^{K} s_k \tag{2.4}$$

The index is bounded between -1 (worst performance) to 1 (best performance), with values close to 0 indicating overlapping clusters (16). The optimal cluster number is achieved by maximizing the index.

– **Davies-Bouldin index**: for each cluster, the similarities between a cluster and all other clusters are computed, with the highest similarity value assigned to that cluster (5). The index is computed by the mean of all clusters' similarities according to (2.5),

$$Davies - Bouldin\ index = \frac{1}{K}\sum_{k=1}^{K} \max_{k' \neq k} \left(\frac{\delta_k + \delta_{k'}}{\triangle_{kk'}} \right) \tag{2.5}$$

whit K the total number of clusters, δ_k the mean distance between the k cluster's points to the k cluster's centroid, $\delta_{k'}$ the mean distance between the k' cluster's points to the k' cluster's centroid, and $\triangle_{kk'}$ the distance between the k and k' clusters' centroids. The index is bounded between 0 (best performance) to $+\infty$ (worst performance) (16), where the optimal cluster number is achieved by minimizing the index.

– **SD index**: for the index calculation, two different properties are considered, the *average scattering for clusters*, and the *total separation between clusters* (6; 11). The *average scattering for clusters*, $\mathscr{S}$, quantifies the compactness based on the variances of the cluster's objects, according to (2.6), with $V^{\{k\}}$ the variance vector of the cluster k.

$$\mathscr{S} = \frac{\frac{1}{K}\sum_{k=1}^{K} \| V^{\{k\}} \|}{\| V \|} \tag{2.6}$$

The *total separation between clusters*, $\mathscr{D}$, quantifies the separation differences based on the distances between clusters' centroids, according to (2.7), with D_{min} and D_{max} the smallest and largest distance between the clusters' centroids, respectively.

$$\mathscr{D} = \frac{D_{max}}{D_{min}} \sum_{k=1}^{K} \frac{1}{\sum_{\substack{k'=1 \\ k' \neq k}}^{K} \| G^{\{k\}} - G^{\{k'\}} \|} \tag{2.7}$$

The index is computed by the sum of the two terms as stated in (2.8), with α a weight equal to the value of $\mathscr{D}$ obtained by the partition with the greatest number of clusters. The optimal cluster number is achieved by minimizing the index.

$$SD\ index = \alpha\mathscr{S} + \mathscr{D} \tag{2.8}$$

– **S_Dbw index**: two different properties are considered, the *average scattering for clusters* and the *between-cluster density* (6; 11). The *average scattering for clusters*, $\mathscr{S}$, was previously defined in (2.6). The *between-cluster density*, $\mathscr{G}$, takes density into account to measure the inter-cluster separation according to (2.9), with $R_{kk'}$ the ratio between the points' density at the midpoint of the clusters' centroids, and the largest density at the two clusters' centroids.

$$\mathscr{G} = \frac{2}{K(K-1)} \sum_{k<k'} R_{kk'} \tag{2.9}$$

The index is computed by the sum of the two terms as stated in (2.10). The optimal cluster number is achieved by minimizing the index.

$$S\_Dbw\ index = \mathscr{S} + \mathscr{G} \tag{2.10}$$

- **External clustering validation**: considered when ground truth labels are available (1). The presented validation metrics are currently used in the OD context when labels are available, being bounded between 0 to 1 (or 0% to 100%).

  – **Recall**: also know as True Positive Rate (TPR), is the ratio between the *True Positives* (true outliers) that the model correctly identified, and the sum of the *True Positives* and *False Negatives* (missed outliers by the model), according to (2.11). A good OD model should maximize this ratio.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2.11}$$

  – **False Positive Rate (FPR)**: is the ratio between the *False Positives* (false outliers) that the model wrongly identified, and the sum of the *False Positives* and *True Negatives* (true inliers), according to (2.12). A good OD model should minimize this ratio.

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives + True\ Negatives} \tag{2.12}$$

  – **F_β Score**: measures the overall effectiveness of the OD model (2) by the harmonic mean of the *Precision* and *Recall*, according to (2.13), with β a real positive number such that, *Recall* is considered β times as important as *Precision*. A good OD model should maximize this measure.

$$F_\beta = \left(1+\beta^2\right) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} \tag{2.13}$$

  – **Area Under the Receiver Operating Characteristic Curve (ROC AUC)**: the Receiver Operating Characteristic (ROC) curve is a visual analysis of the trade-off between TPR and FPR, by plotting a two-dimensional graph with the FPR on the x-axis

and the TPR on the y-axis (2). The edge points of the ROC curve are located at coordinates $(0,0)$ and $(100,0)$, where a random method will have its performance along a diagonal line connecting these two points (1). The ROC AUC is the area below the ROC curve.

Chapter 3

Data Processing

This chapter describes and justifies all the data processing phases. The data acquisition details and the lithography process attributes are presented in Section 3.1. The data preparation methodology is described and fully justified in Section 3.2.

3.1 Data Acquisition

The data used for this work was gathered by the Company during the past twenty months (from June 2021 to March 2023), where a group of lithography process attributes were collected from production lines L4 and L8. Data collection occurred along several mass production events with a time step acquisition of 15 minutes between instances. The multivariate time-series were continuously stored on a Structured Query Language (SQL) database, which was then available for query. No labels were held in the database or provided during this work, as well as no notes or remarks from any mass production event (production line start or stoppage).

The name of the process attributes of interest, description, format and unit metric are presented in Table 3.1 and Table 3.2, concerning production lines L4 and L8, respectively. Note that other process and performance attributes were available in the database. Nevertheless, the ones selected and presented in the tables were the only ones that matched the attributes' selection requirements concerning scope, number of missing values and zero values. At the end of this step, we had a total of ten attributes (for each production line dataset) allocated to our lithography process.

3.2 Data Preparation

Based on an initial Exploratory Data Analysis (EDA) performed and due to specific characteristics of the datasets, several data processing challenges were found. These challenges motivated the data processing methods used in this work. The methods used will be detailed in the following sub-sections, and their assumptions will be justified. Note that the adopted approaches aimed

Table 3.1: Lithography process attributes of interest collected from production line L4.

Process Attribute (ID)	**Description**	**Format**	**Unit Metric**
Timestamp	Timestamp when each attribute was acquired.	Datetime	yyyy-mm-dd hh:mm:ss
Real Gas Consumption (10)	Dryer (Oven) gas consumption.	Float	Not specified
CE1 Total Active Energy (1)	Active energy delivered (into the load), and the active energy received (out from the load).		
CE2 Feeder Active Energy (5)	Active energy delivered (into the load).		
CE1 Reactive Energy (2)	Reactive energy delivered (into the load), and the reactive energy received (out from the load).		
CE2 Reactive Energy (8)			
CE1 Apparent Energy (6)	Combination between active energy and reactive energy.		
CE2 Apparent Energy (3)			
Entrance Temperature (7)	Dryer (Oven) temperatures.		Degrees Celsius (ºC)
Middle Temperature (9)			
Exit Temperature (4)			

Table 3.2: Lithography process attributes of interest collected from production line L8.

Process Attribute (ID)	**Description**	**Format**	**Unit Metric**
Timestamp	Timestamp when each attribute was acquired.	Datetime	yyyy-mm-dd hh:mm:ss
Real Gas Consumption (10)	Dryer (Oven) gas consumption.	Float	Not specified
Velocity Record (7)	Conveyor belt velocity.		
Active Energy (8)	Active energy delivered (into the load).		
Reactive Energy (2)	Reactive energy delivered (into the load), and the reactive energy received (out from the load).		
Apparent Energy (3)	Combination between active energy and reactive energy.		
Entrance Temperature (1)	Dryer (Oven) temperatures.		Degrees Celsius (ºC)
Middle Temperature 1 (6)			
Middle Temperature 2 (5)			
Exit Temperature (9)			
Incinerator Temperature (4)			

to maximize the dataset quality, as well as the instances quantity. The data processing workflow applied to the datasets and the related data loss is quantified at the end of this chapter in Table 3.5.

3.2.1 Initial Processing

The data processing started with a standard approach where duplicated values, zero values and non-time step multiples were removed.

Duplicated values were similar instances that presented the same attribute values for the same timestamp. Only one instance was considered valid for each duplicated group of instances, and the remaining were removed.

Zero values were instances that presented zero values for all the attributes in the same timestamp. These instances were removed from the datasets since no helpful information could be extracted.

Non-time step multiples were instances where the timestamp was not a fifteen-minute multiple. These instances were also removed from the datasets.

3.2.2 Outliers

Two different outlier categories were identified in the datasets: standard outliers and extreme outliers. Although this work aims to build an OD model capable of cataloguing anomalous patterns, feeding outliers directly to the model training will degrade the model's performance. Therefore, outliers must be identified and removed.

The root cause of standard outliers was not specified. Nevertheless, the extreme outliers occurrence might be related to the production line start-up, where the equipment requests a high amount of resources (e.g., energy), and consumption spikes will be formed (extreme outliers). The possibility of data error due to equipment or data infrastructure malfunction is not excluded.

To address the standard outliers, an empirical threshold was considered based on the time-series graphical evaluation during the initial EDA. Concerning extreme outliers, Tukey's method was used considering a threshold of three Interquartile Range (IQR) from its quantile Q_1 and Q_3. Tables 3.3 and 3.4 present the standard and extreme outliers removal considerations for production lines L4 and L8 datasets.

Table 3.3: Standard and extreme outliers removal considerations adopted for production line L4 dataset.

Process Attribute	**Outlier Removal Consideration**
Real Gas Consumption	Values above the maximum extreme outlier limits were removed.
CE1 Total Active Energy	
CE2 Feeder Active Energy	
CE1 Reactive Energy	
CE2 Reactive Energy	
CE1 Apparent Energy	
CE2 Apparent Energy	
Entrance Temperature	Values above 300°C were removed.
Middle Temperature	
Exit Temperature	

Table 3.4: Standard and extreme outliers removal considerations adopted for production line L8 dataset.

Process Attribute	**Outlier Removal Consideration**
Real Gas Consumption	Values above 45 were removed.
Velocity Record	Values above 8000 were removed.
Active Energy	Values above the maximum extreme outlier limits were removed.
Reactive Energy	
Apparent Energy	
Entrance Temperature	No outliers were removed.
Middle Temperature 1	
Middle Temperature 2	
Exit Temperature	
Incinerator Temperature	

3.2.3 Missing Values

Missing values were detected in all attributes in different degrees of severity. This severity was quantified, and several approaches were used, from simply removing the instance to a more complex approach involving interpolation between neighbour instances to replace the missing ones. The root cause of missing values was not specified. Nevertheless, their occurrence might be related to shutdown periods (bank holidays, weekends and holidays) when no production is performed. However, a connection between sensors and the database might still be established, resulting in the occurrence of missing values.

To quantify the number of missing values per attribute, *continuous data reception blocks* were computed. We defined a *continuous data reception block* as a group of continuously collected instances based on a specified time step (15 minutes in our case). This approach was adopted to find blocks of continuously collected data to identify and quantify the start and stop of mass production events.

The number of missing values was then quantified in percentage by the number of missing values in the block per total number of instances in the same block. The missing values incidence maps for the production lines L4 and L8 datasets are presented in Figures 3.1 and 3.2, respectively. Each percentage range was colour-coded from green (low missing values incidence) to red (high missing values incidence). Single missing values were assigned with a red cross, and weekends were colour-coded in light grey. Subsequently, some conclusions are highlighted from the presented Figure 3.1 and Figure 3.2:

- Although the majority of the maps present a low missing values incidence (green *continuous data reception blocks*), two large time frames with a high missing values incidence (red *continuous data reception blocks*) are observed on the left and right of both figures. The root cause for these occurrences was not specified. Nevertheless, the one on the left might be related to the plant shutdown for the summer holidays;

- For the production line L8 dataset, a degradation of the data is noticeable (yellow *continuous data reception blocks*) from 2022-08 to 2022-09 for all the attributes of interest, whit the exception of the *Active Energy* attribute;

- For the production line L8 dataset, a degradation of the data is noticeable (orange and yellow time frames) for the *Velocity Record* attribute at specific time frames of the dataset.

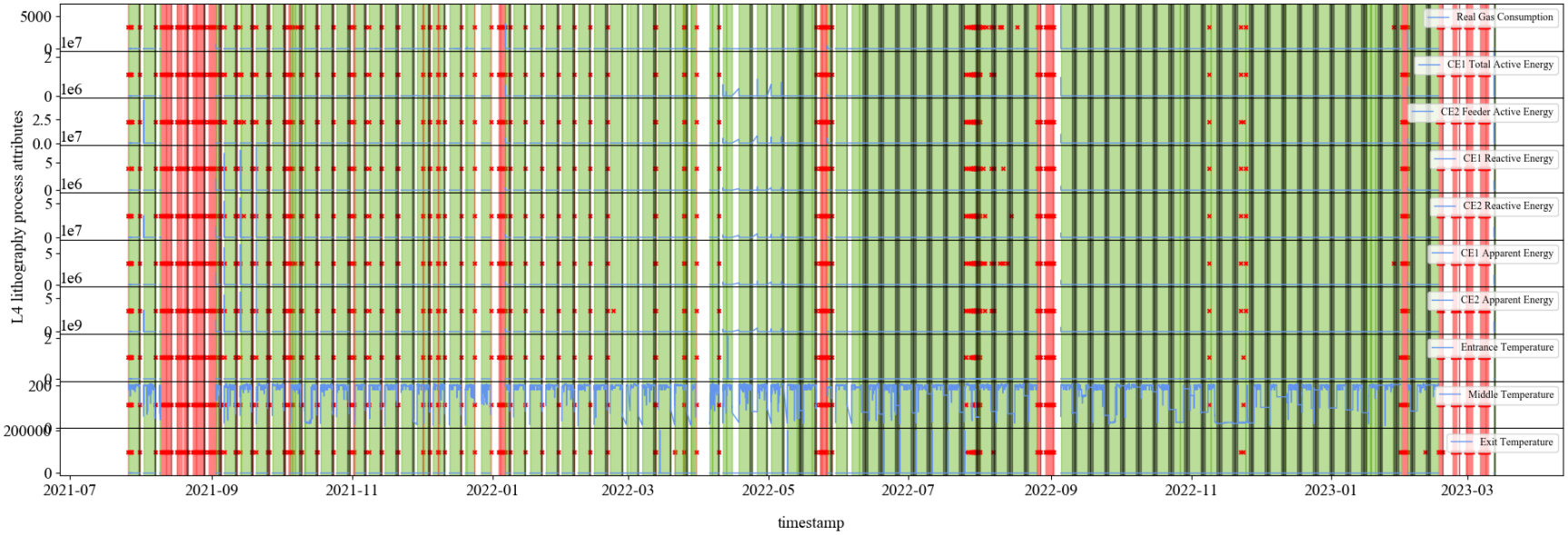

Figure 3.1: Incidence map of missing values for the attributes of interest in production line L4 dataset.

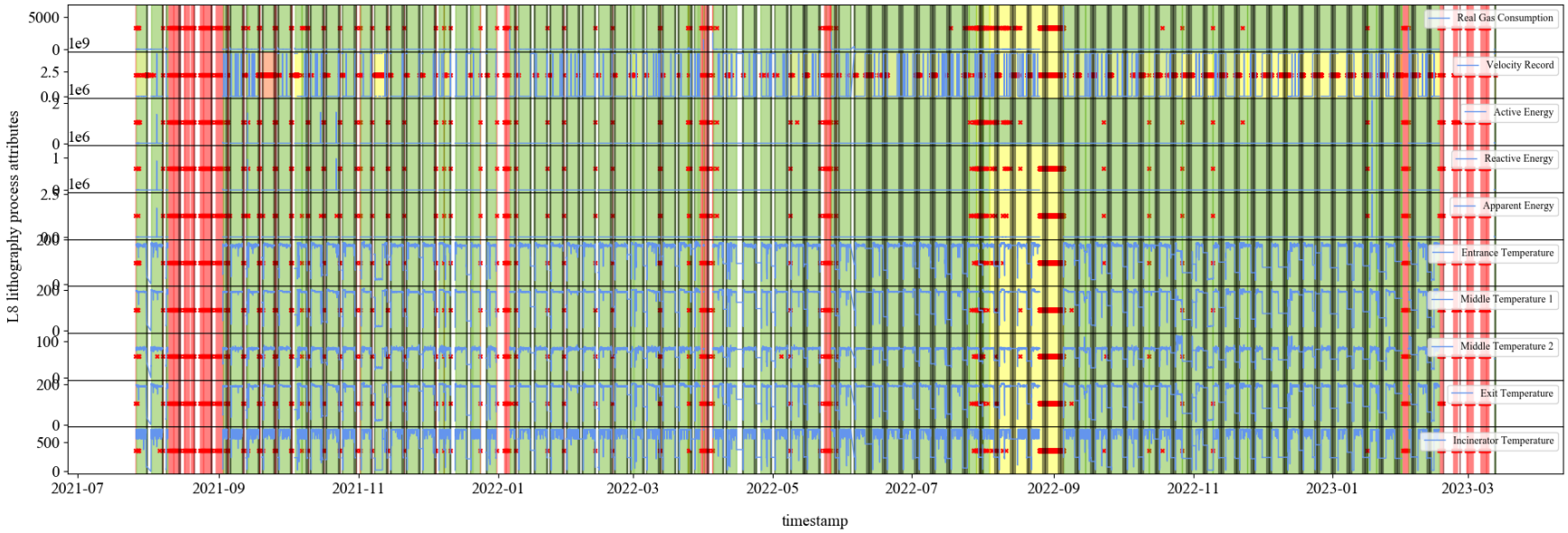

Figure 3.2: Incidence map of missing values for the attributes of interest in production line L8 dataset.

Based on these findings, instances in which all the attribute values were missing were removed from the datasets in the first place.

Secondly, instances with at least five missing values (out of ten possible for each dataset) were removed from the datasets. This line of thinking is based on the principle that an instance with at least 50% of its expected information missing will not bring additional insight into the problem. The initial distribution of missing values per instance and attribute is presented in Figure 3.3 and 3.4 for the production lines L4 and L8 datasets, respectively.

Based on both figures, we concluded that for the production line L4 dataset, most instances have only one attribute value missing. The same situation occurs for the production line L8, together with instances that have nine attribute values missing.

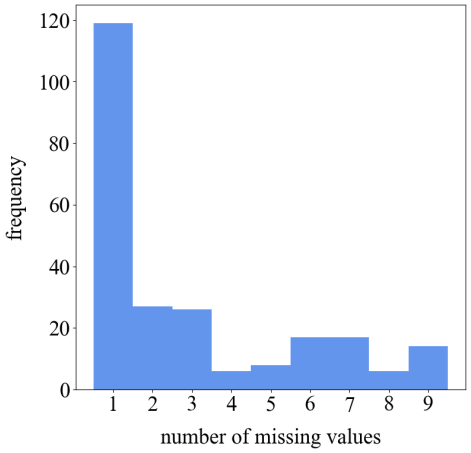

Figure 3.3: Missing values distribution per instance for production line L4 dataset.

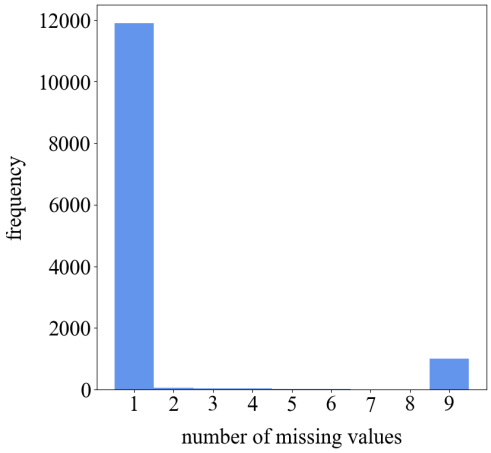

Figure 3.4: Missing values distribution per instance for production line L8 dataset.

Thirdly, instances with more than four consecutive missing values were removed from the datasets. Based on the established time step (15 minutes), four consecutive missing values result in one hour of data loss, making any interpolation unrealistic. The distribution of the number of consecutive missing values per instance and attribute is presented in Figure 3.5 and Figure 3.6.

Based on both figures, we concluded that for the production line L4 dataset, most of the consecutive missing value occurrences comprise only one instance (a total of 15 minutes of data interruption). Concerning the production line L8, most consecutive missing value occurrences occur in the range between 1 to 99. Note that, in the range between 900 and 999, some consecutive missing value occurrences are also noticeable, pointing out that some attribute values were not collected during extensive periods.

3.2.4 Repeated Values

Repeated values were detected when a sensor recorded the same value during an extended period. Although the root cause of repeated values was not specified, different scenarios could justify this phenomenon. Shutdown periods when the production line was stopped, but the sensors

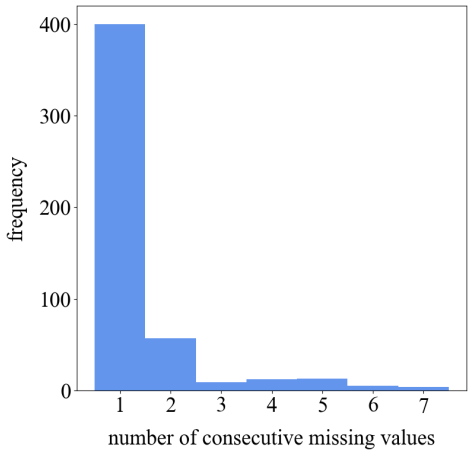

Figure 3.5: Consecutive missing values distribution per instance for production line L4 dataset.

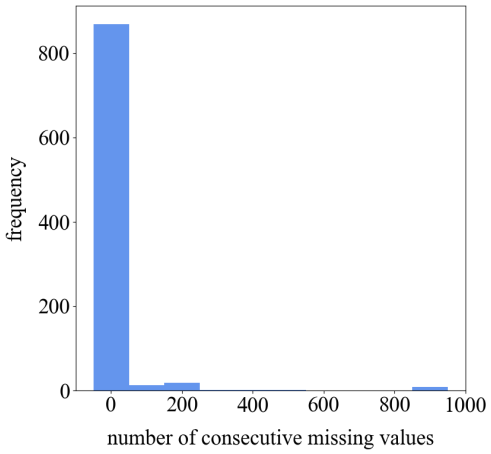

Figure 3.6: Consecutive missing values distribution per instance for production line L8 dataset.

were still uploading data to the database. Sensors malfunction with the last same reading being continuously updated to the database. Repeated values were indirectly removed simultaneously as the previous consecutive missing values.

3.2.5 Isolated Values

Isolated values were defined as short *continuous data reception blocks*, which remained after applying all data processing methods. *Continuous data reception block* with less than eight instances (equivalent to two hours of data) were removed.

Table 3.5 presents the applied data processing workflow and related data loss quantification. Concerning the production line L4 dataset, approximately 10% of the data was lost, mainly due to the removal of duplicated instances (step one of the workflow). Regarding the production line L8 dataset, around 30% of the data was lost, mainly due to consecutive missing values instance removal (step seven of the workflow). The number of missing values in the *Velocity Record* attribute was the main contributor to the data loss, confirmed by the incidence map of missing values presented in Figure 3.2. Note that for step nine of the workflow, no decrease in the data was noticed since interpolation was part of the data processing phase but had no contribution to data loss.

All the challenges were overcome at the end of the data processing phase. Figure 3.7 and Figure 3.8 present the final missing values incidence maps for the production lines L4 and L8 datasets, respectively. Based on this, we concluded that the data processing phase was finished, and the datasets were ready for the next step, data modelling.

Table 3.5: Number of instances lost during data processing for production lines L4 and L8 datasets.

		Production Line	
Workflow	**Data Processing Method**	L4	L8
	Total number of instances	**46,974 (100.00%)**	**47,043 (100.00%)**
1	Remove duplicated instances	42,424 (90.31%)	45,527 (96.78%)
2	Remove instances in which all attribute values were zero	42,412 (90.29%)	45,522 (96.77%)
3	Remove non-time step multiples	42,406 (90.28%)	45,519 (96.76%)
4	Remove instances in which all attribute values were missing	42,191 (89.82%)	45,340 (96.38%)
5	Remove standard and extreme outliers values	42,191 (89.82%)	45,340 (96.38%)
6	Remove instances with more than 5 missing values per attribute	42,137 (89.70%)	44,309 (94.19%)
7	Remove instances with more than 4 consecutive missing values	42,113 (89.65%)	32,835 (69.80%)
8	Remove isolated values	42,082 (89.59%)	32,737 (69.59%)
9	Interpolation	42,082 (89.59%)	32,737 (69.59%)

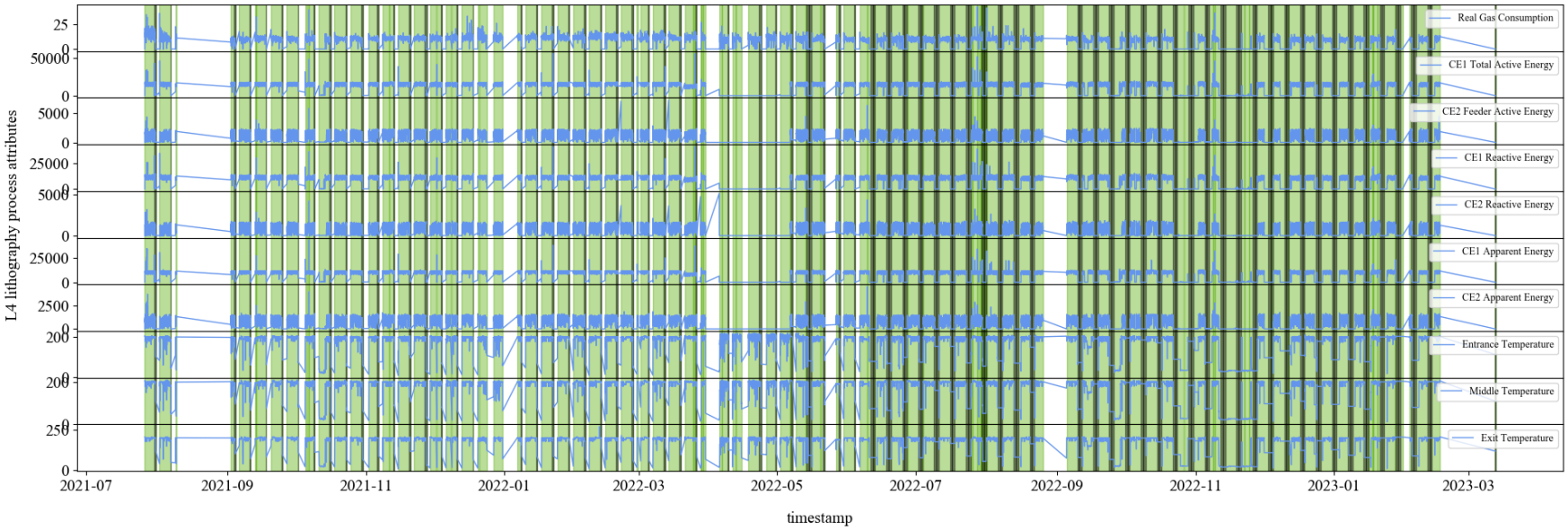

Figure 3.7: Incidence map of missing values for the attributes of interest in production line L4 dataset at the end of the data processing phase.

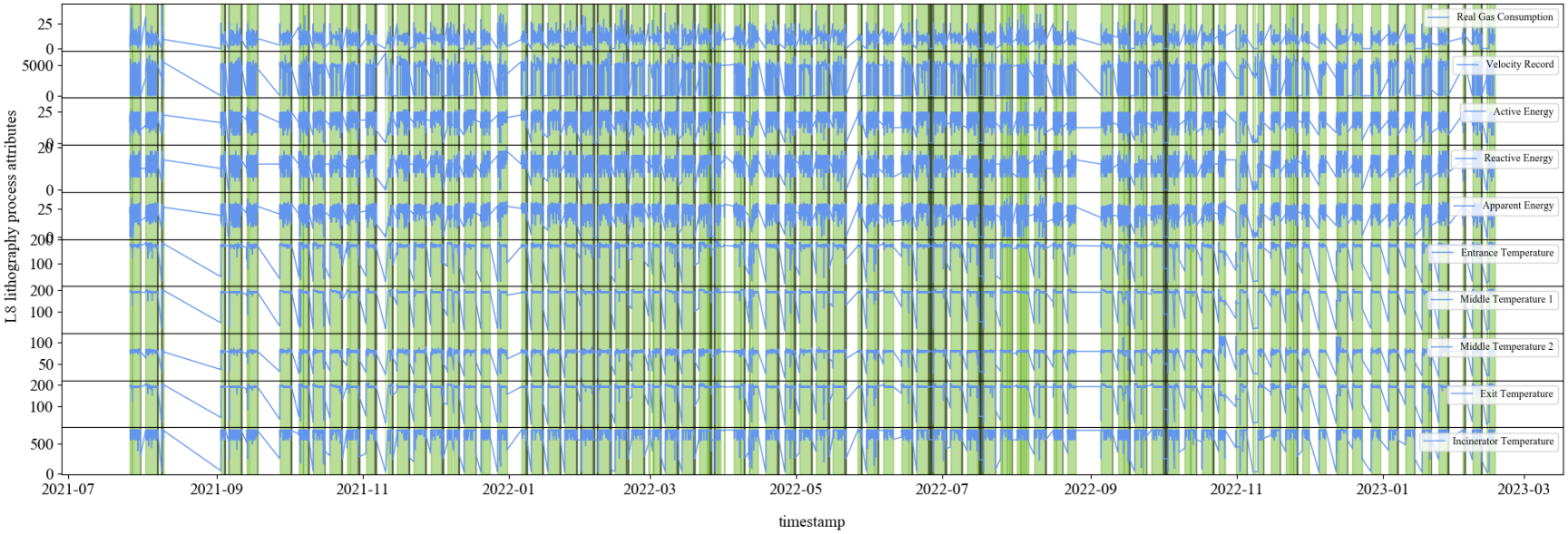

Figure 3.8: Incidence map of missing values for the attributes of interest in production line L8 dataset at the end of the data processing phase.

Chapter 4

Data Modelling

This chapter describes and justifies all the data modelling methods used in this work. The approach adopted for the datasets' dimensionality reduction will be described in Section 4.1. Different clustering methods were used since no labels were held in the database or provided during this work. A detailed explanation of how the clustering algorithms were used is described in Section 4.2. The most promising clustering models were selected, and the originated clusters were statistically characterized to map the lithography process behaviour fully. At the end of the clustering modelling, the instances of each dataset were labelled. The obtained results are presented in Section 4.3. After that, different OD algorithms were tested to learn what was considered a normal lithography process behaviour. The results of these implementations are described in Section 4.4.

4.1 Principal Components Analysis

As is well known, datasets with a large number of attributes (higher dimensionalities) could be challenging to handle for clustering algorithms (*i.e.*, the curse of dimensionality problem). Clustering algorithms use distance measures (Euclidean, Minkowski or others) to quantify the similarity between instances. The existence of a large number of attributes forces every instance to appear equally spaced between each other. Therefore, all the instances might appear equally distanced, resulting in no meaningful insights retrieved from the clustering analysis.

Since ten attributes were available for each production line dataset, a dimensionality reduction approach based on PCA was applied to reduce the problem dimensionality and additionally to support a visual self-evaluation of the models. A visual self-evaluation can only be performed when a dimensionality reduction to two or three dimensions is achieved. Although we understand that PCA should not be considered a visualization tool, the fact that it reduces the problem dimensionality will certainly assist us during the cluster identification.

To define the number of principal components to be selected, a variance explainability threshold of 90% was considered. The variance explainability for each principal component and the

graphical representation of the selected principal components are presented in Figure 4.1 and Figure 4.2, as well as in Figure 4.3 and Figure 4.4 for the production lines L4 and L8 datasets, respectively.

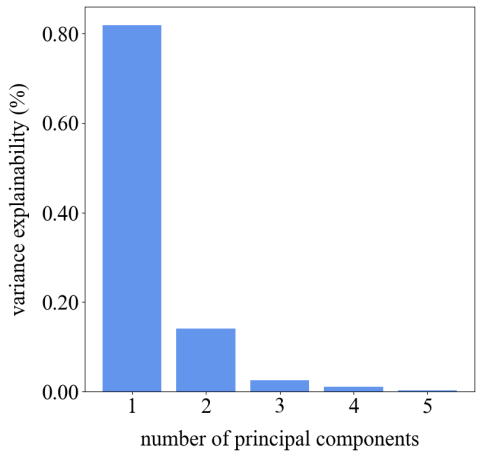

Figure 4.1: Variance explainability of the principal components for production line L4 dataset. Together, the principal components 1 (81.88%) and 2 (14.05%) explain 95.93% of the variance.

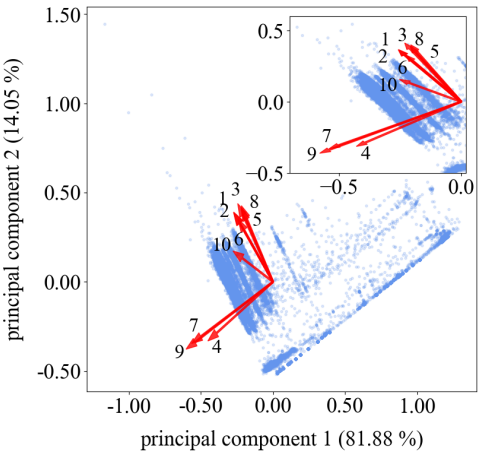

Figure 4.2: Graphical representations of the principal components for production line L4 dataset, with the related variance explainability between brackets. Attributes' IDs available in Table 3.1.

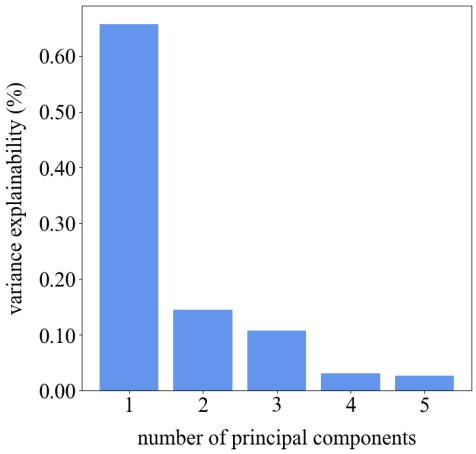

Figure 4.3: Variance explainability of the principal components for production line L8 dataset. Together, the principal components 1 (65.73%), 2 (14.49%), and 3 (10.79%) explain 91.01% of the variance.

The importance of each attribute in each principal component of interest is also presented in Figure 4.2 and Figure 4.4. The importance of each attribute is reflected by the magnitude of the eigenvectors (red arrows), where a higher magnitude is related to a higher attribute importance. The three most relevant attributes for each principal component are presented in Table 4.1. Based

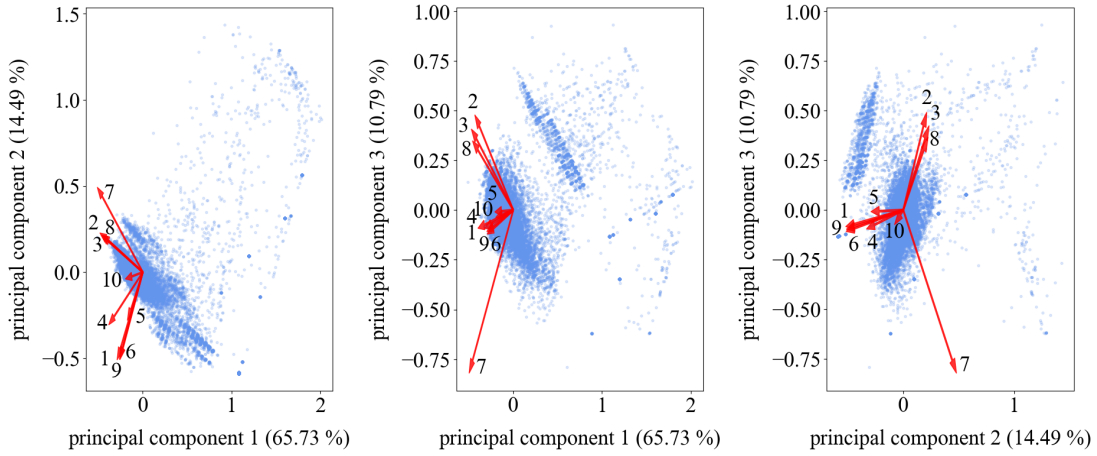

Figure 4.4: Graphical representations of the principal components for production line L8 dataset, with the related variance explainability between brackets. Attributes' IDs available in Table 3.2.

on the presented results for production line L4, we concluded that the most relevant process attributes are related to temperatures and energies. The relevant process attributes for production line L8 are associated with velocity, energies and temperatures. Conversely, the *Real Gas Consumption* process attribute is not considered relevant for any of the principal components for both production lines.

Note that the timestamp was not included as an attribute in the presented dimensionality reduction approach. Additional investigations were conducted where this attribute was included. Nevertheless, it led to a higher dimensionality model (a higher number of principal components to consider), making it more complex to apply and difficult to visually self-evaluate during the clustering phase.

At the end of the dimensionality reduction phase, it was possible to reduce the production line L4 dataset from ten to two dimensions, with a total variance explainability of 95.93%. This degree of reduction is justified by the correlation observed between two major groups of attributes. In

Table 4.1: Top three most relevant attributes for each principal component, for production line L4 and L8 datasets.

	Attribute Name (Eigenvector Magnitude)	
Principal Component Number	**Production Line L4**	**Production Line L8**
1	*Middle Temperature* (0.540)	*Velocity Record* (0.460)
	Entrance Temperature (0.501)	*Apparent Energy* (0.417)
	Exit Temperature (0.392)	*Active Energy* (0.401)
2	*CE2 Apparent Energy* (0.374)	*Entrance Temperature* (0.444)
	CE2 Reactive Energy (0.360)	*Middle Temperature 1* (0.442)
	CE2 Feeder Active Energy (0.349)	*Velocity Record* (0.441)
3	-	*Velocity Record* (0.757)
		Reactive Energy (0.424)
		Apparent Energy (0.359)

PCA, correlation is observed when a group of eigenvectors (red arrows) are pointing in the same direction. Based on Figure 4.2, a correlation between temperature attributes (IDs 4, 7 and 9) is observed, as well as between energy attributes (IDs 1, 2, 3, 5, 6 and 8). Based on this, it was possible to resume the production line L4 dataset into two principal components.

Concerning the production line L8, the dataset dimension was reduced from ten to three dimensions, with a total variance explainability of 91.01%. Based on Figure 4.4, a correlation between energy attributes (IDs 2, 3 and 8), together with temperature attributes (IDs 1, 4, 5, 6 and 9), is observed. In addition to the *Velocity Record* attribute (ID 7), it was possible to resume the production line L8 dataset into three principal components.

Additionally, Figure 4.2 and Figure 4.4 show that after PCA application, both datasets were split into different clusters. In Section 4.2, several clustering methods were used to detect significant clusters, followed by their statistical characterization and association with the problem.

4.2 Clustering

Clustering methods were applied in this work since no labels or annotations from the lithography process were available. These unsupervised learning methods were then used to detect significant clusters in the two datasets after PCA application.

Supported on the *Scikit-Learn* library (16), four different clustering algorithms were applied, namely Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) and Mean Shift (MS). Other algorithms were also tested nevertheless, due to the lack of computation power, no results were possible to obtain. Based on a hyperparameter grid-search, different models were created from each clustering algorithm.

To find the best model, a total of 24,602 models were generated. Their performance was evaluated according to the metrics described in Section 2.6 and based on visual self-evaluation (the expected clusters' visual separation). To improve the visual self-evaluation, transparency was applied in the cluster visualisation. Therefore, the clusters' density variations are more noticeable and easy to self-evaluate visually. The performance of the best models achieved for each number of clusters is presented in Table 4.2. Concerning the GMM algorithm, a mean and standard deviation value is presented due to the random seed given to the method to initialise the parameters (a total of 30 random seeds were used for each GMM model).

To reduce the problem analysis complexity, the number of studied clusters was constrained between two to ten. A minimum of two clusters was considered since a production process should have at least two operating modes. A maximum of ten clusters were established to delimit and simplify the problem analysis. Additionally, other information of interest was also encoded in the table, namely:

- The visual self-evaluation was divided into two categories: visual overfitted (o) and visual underfitted (u) models. An example of a visual overfitted model is presented in Figure

4.5, which is considered a more complex model than the expected cluster visual separation. Conversely, an example of a visual underfitted model is presented in Figure 4.6, which considers a simpler model than the expected cluster visual separation;

- Table cells highlighted in bold indicate the best visual self-evaluated models;
- Table cells coloured in grey indicate evaluation metrics from models obtained by the same hyperparameters.

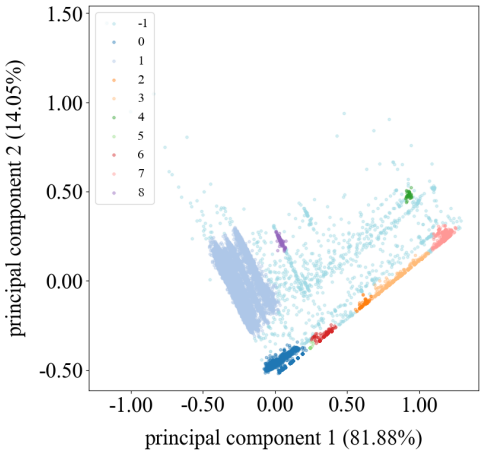

Figure 4.5: Example of a visual overfitted DBSCAN model, with the lower region subdivided into multiple clusters.

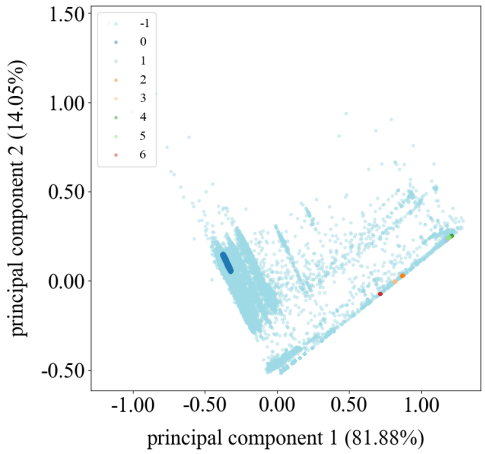

Figure 4.6: Example of a visual underfitted DBSCAN model, with the clustering not properly fitting the dataset.

Throughout the models' evaluation process, it was possible to conclude that *SD Index* was the evaluation metric that better reproduced the more expected visual self-evaluation results. Therefore, the *SD Index* was considered a guidance metric during the best model selection.

Based on the results presented in Table 4.2 concerning the production line L4 dataset, no satisfactory models were possible to obtain when GMM and MS algorithms were applied since all the obtained models were considered visually underfitted. Conversely, using DBSCAN and HDBSCAN algorithms produced three and four suitable models, respectively. Therefore, a five-cluster model obtained by the DBSCAN algorithm was considered the best result achieved, based on its visual self-evaluation, lowest *SD Index* (5.445) and model complexity. The final clustering result for the selected model is presented in Figure 4.7.

Regarding the production line L8 dataset, no acceptable models were obtained when DBSCAN and MS algorithms were applied since, for both cases, the obtained models were considered visually overfitted or underfitted. Contrarily, using GMM and HDBSCAN algorithms produced one and three suitable models, respectively. Consequently, a four-cluster model obtained by the HDBSCAN algorithm was considered the best result achieved, based on its visual self-evaluation, lowest *SD Index* (3.089) and model complexity. The final clustering result for the selected model is presented in Figure 4.8.

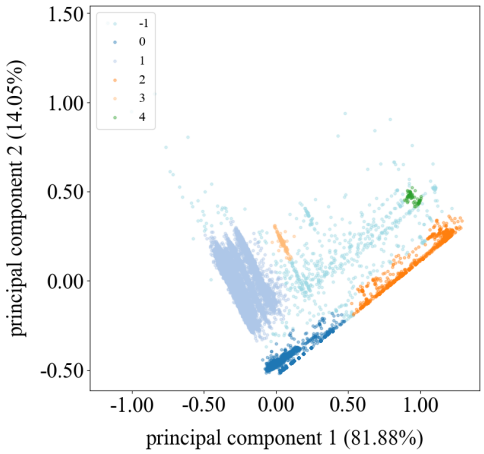

Figure 4.7: Best DBSCAN clustering model of the principal components for production line L4 dataset. The model was obtained with an *epsilon* of 0.06, a *minimum number of samples* in the neighbourhood of 65, and considering a *Euclidean* distance. Instances assigned to the -1 cluster are considered noise.

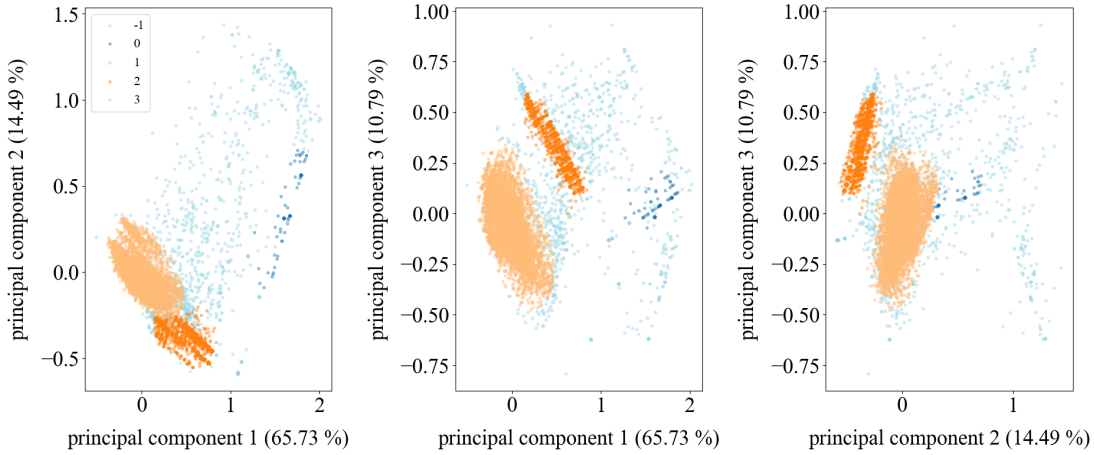

Figure 4.8: Best HDBSCAN clustering model of the principal components for production line L8 dataset. The model was obtained with a *cluster epsilon* of 0.005, a *minimum number of samples* in the neighbourhood of 405, a *minimum cluster size* of 100, and considering a *Euclidean* distance. Instances assigned to the -1 cluster are considered noise.

At this phase, the processed data was already split into clusters. In the next Section, each cluster will be statistically characterized to identify specific operating modes and comprehensively map the lithography process behaviour.

Table 4.2: Evaluation metrics of the best clustering models for production line L4 and L8 datasets.

		Evaluation Metrics (Visual Self-Evaluation)							
		Production Line L4				Production Line L8			
Clustering Algorithm	**Number of Clusters**	**Silhouette Index**	**Davies-Bouldin Index**	**SD Index**	**S_Dbw Index**	**Silhouette Index**	**Davies-Bouldin Index**	**SD Index**	**S_Dbw Index**
DBSCAN	2	0.994 (u)	0.010 (u)	2.192 (u)	0.010 (u)	0.907 (u)	0.068 (u)	1.762 (u)	0.285 (u)
	3	0.994 (u)	0.009 (u)	3.427 (u)	0.005 (u)	0.826 (u)	0.151 (u)	4.023 (u)	0.307 (u)
	4	0.975 (u)	0.024 (u)	4.102 (u)	0.016 (u)	0.909 (u)	0.058 (u)	4.210 (u)	0.089 (u)
	5	0.970 (u)	0.045 (u)	**5.445 (s)**	0.026 (u)	0.906 (u)	0.061 (u)	4.145 (u)	0.071 (u)
	6	0.956 (u)	0.091 (u)	**6.091**	0.031 (u)	0.853 (u)	0.088 (u)	6.301 (u)	0.080 (u)
	7	0.930 (u)	0.120 (u)	**7.292**	0.098 (u)	0.845 (u)	0.114 (u)	6.428 (u)	0.091 (u)
	8	0.925 (u)	0.168 (u)	9.983 (o)	0.112 (u)	0.837 (u)	0.137 (u)	7.252 (o)	0.077 (u)
	9	0.915 (u)	0.168 (u)	16.143 (o)	0.122 (u)	0.788 (u)	0.147 (u)	7.114 (o)	0.062 (u)
	10	0.907 (u)	0.160 (u)	14.774 (o)	0.074 (u)	0.784 (u)	0.199 (u)	12.592 (o)	0.116 (u)
GMM	2	0.670 ± 0.068 (u)	0.557 ± 0.166 (u)	2.647 ± 0.376 (u)	0.853 ± 0.285 (u)	0.690 ± 0.047 (u)	1.003 ± 0.125 (u)	3.801 ± 0.430 (u)	1.873 ± 0.292 (u)
	3	0.644 ± 0.051 (u)	0.793 ± 0.454 (u)	4.625 ± 0.808 (u)	0.794 ± 0.226 (u)	**0.590 ± 0.137**	**0.973 ± 0.201**	**6.567 ± 2.780**	**1.348 ± 0.228**
	4	0.617 ± 0.091 (u)	0.672 ± 0.390 (u)	7.628 ± 4.261 (u)	0.551 ± 0.133 (u)	0.475 ± 0.177 (o)	1.161 ± 0.294 (o)	8.584 ± 3.101 (o)	1.148 ± 0.105 (u)
	5	0.483 ± 0.168 (u)	1.022 ± 0.775 (u)	12.324 ± 5.654 (u)	0.515 ± 0.192 (u)	0.448 ± 0.167 (o)	1.140 ± 0.296 (o)	10.453 ± 4.778 (o)	1.015 ± 0.097 (o)
	6	0.526 ± 0.103 (u)	0.802 ± 0.373 (u)	12.408 ± 4.928 (u)	0.440 ± 0.163 (u)	0.379 ± 0.163 (o)	1.236 ± 0.471 (o)	14.499 ± 12.254 (o)	0.948 ± 0.106 (o)
	7	0.455 ± 0.115 (u)	0.898 ± 0.358 (u)	21.354 ± 6.866 (u)	0.447 ± 0.134 (u)	0.333 ± 0.165 (o)	1.390 ± 0.652 (o)	20.159 ± 19.013 (o)	0.852 ± 0.130 (o)
	8	0.464 ± 0.086 (u)	0.889 ± 0.261 (u)	21.841 ± 6.172 (u)	0.407 ± 0.110 (u)	0.317 ± 0.163 (o)	1.561 ± 1.085 (o)	24.943 ± 25.342 (o)	0.787 ± 0.124 (o)
	9	0.449 ± 0.067 (u)	0.909 ± 0.249 (u)	25.159 ± 8.218 (u)	0.396 ± 0.081 (u)	0.276 ± 0.165 (o)	1.719 ± 1.099 (o)	33.912 ± 30.585 (o)	0.716 ± 0.122 (o)
	10	0.444 ± 0.070 (u)	0.960 ± 0.250 (u)	28.056 ± 9.141 (u)	0.373 ± 0.068 (u)	0.276 ± 0.172 (o)	1.832 ± 1.373 (o)	37.615 ± 35.623 (o)	0.681 ± 0.105 (o)
HDBSCAN	2	0.675 (u)	0.382 (u)	2.165 (u)	0.604 (u)	0.771 (u)	0.349 (u)	3.191 (u)	0.801 (u)
	3	0.730 (u)	0.384 (u)	3.496 (u)	0.408 (u)	0.772 (u)	0.204 (u)	2.880 (u)	0.360 (u)
	4	0.745 (u)	0.328 (u)	4.063 (u)	0.340 (u)	0.773 (u)	0.173 (u)	**3.089 (s)**	0.253 (u)
	5	0.737 (u)	0.268 (u)	5.472 (u)	0.214 (u)	0.775 (u)	0.162 (u)	3.405 (u)	0.204 (u)
	6	0.728 (u)	0.280 (u)	6.041 (u)	0.187 (u)	0.775 (u)	0.173 (u)	4.611 (u)	0.168 (u)
	7	0.706 (u)	0.287 (u)	**7.901**	0.178 (u)	**0.758**	0.281 (o)	**5.169**	**0.235**
	8	0.703 (u)	0.295 (u)	8.510 (u)	**0.172**	0.759 (o)	0.273 (o)	6.036 (o)	0.207 (o)
	9	0.709 (u)	0.281 (u)	**10.401**	0.160 (u)	0.758 (o)	0.254 (o)	6.877 (o)	0.189 (o)
	10	0.698 (u)	0.294 (u)	**12.932**	0.160 (u)	0.758 (o)	0.270 (o)	7.963 (o)	0.179 (o)
MS	2	0.844 (u)	0.271 (u)	2.002 (u)	0.367 (u)	0.878 (u)	0.285 (u)	2.060 (u)	0.901 (u)
	3	0.862 (u)	0.191 (u)	1.933 (u)	0.219 (u)	-	-	-	-
	4	0.733 (u)	0.367 (u)	3.487 (u)	0.384 (u)	-	-	-	-
	5	-	-	-	-	0.763 (u)	0.522 (u)	3.762 (u)	0.683 (u)
	6	0.711 (u)	0.431 (u)	4.480 (u)	0.436 (u)	-	-	-	-
	10	-	-	-	-	0.798 (o)	0.425 (o)	5.672 (o)	0.370 (o)

(o) - Visual overfitted model
(s) - Selected model
(u) - Visual underfitted model

4.3 Lithography Process Mapping

After concluding the clustering modelling, the production lines L4 and L8 datasets were split into different process operating modes. Each attribute's distribution in each cluster was analysed to identify each operating mode.

Firstly, the distribution of each attribute was tested against the normal distribution using the *Kolmogorov-Smirnov* test (20). This approach was adopted to identify the nature of the statistical distribution of each attribute (parametric or non-parametric) and select the most suitable two independent samples test to compare the same attribute in different clusters.

Secondly, each cluster's attributes were tested against the NPB cluster to search for a significant statistical difference. The non-parametric *Mann-Whitney U* test (20) on two independent samples was adopted to test the difference between attributes' distributions.

This statistical characterization was performed to ensure the correct selection of the NPB cluster and confirm that both clusters were statistically different. For the deployment of both *Kolmogorov-Smirnov* and *Mann-Whitney U* tests, a significance level of 0.05 was applied.

Apart from that, to support the identification of each cluster in the dataset, three main principles were followed:

- The existence of a major cluster representing the NPB. The size of this cluster should always be maximized since the number of instances representing a normal process is always greater than the ones representing an APB;
- The existence of different clusters that represent different process operating modes;
- The smaller clusters are associated with anomalous process behaviours (lithography process outliers).

Figure 4.9 and Figure 4.11 present the attributes' distribution in each cluster. Tables 4.3 and 4.4 show the results obtained for the *Kolmogorov-Smirnov* and *Mann-Whitney U* tests, for the production line L4 and L8 datasets, respectively. The attributes' median and the size of each cluster (number of instances and percentage of the total cluster size) are also presented in the tables.

Concerning the production line L4 dataset, based on the cluster size (66.7% of the total number of instances in the dataset), the attributes' median values (all different from zero), and the time-series presented in Figure 4.10, we selected the cluster 1 as the NPB cluster.

Based on Figure 4.9, on the attributes' median values (all equal to zero, except for the temperature attributes), we concluded that clusters 0 and 2 might be related to the start, stoppage or other interruptions in the mass production process.

Concerning clusters -1, 3 and 4, regarding their cluster size, they represent only 1.6%, 0.3% and 0.2% of the total number of instances in the dataset, respectively. Based on their attributes' median values, and since they are statistically different from the NPB cluster (except for the *CE1*

Total Active Energy attribute in cluster 3), these clusters might fit in the APB category, being considered process outliers.

Concerning the production line L8 dataset, based on the cluster size (89.9% of the total number of instances in the dataset), the attributes' median values (all different from zero), and the time-series presented in Figure 4.12, we selected the cluster 3 as the NPB cluster.

Based on Figure 4.11, on the *Velocity Record* attribute's median value (with two orders of magnitude lower than the NPB), the temperature attributes' median values (similar to the NPB cluster, although with statistically different distributions), we concluded that cluster 1 and 2 might be related to the start, stoppage or other interruptions in the mass production process. Cluster 1 might be considered a particular case of cluster 2 with all the energy and gas consumption attributes' median values equal to zero.

Concerning clusters -1 and 0, based on their cluster size, they represent only 4.1% and 1.4% of the total number of instances in the dataset, respectively. Concerning their attributes' median values, and since they are statistically different from the NPB cluster, these clusters might fit in the APB category, being considered process outliers.

Additional insights related to these different lithography process operating modes must be obtained with the proper supervision of a process engineer.

Although Figure 4.10 and Figure 4.12 present instances grouped in the NPB clusters, some outliers that were incorrectly included are also observed. To the best of our knowledge, these were the best-obtained results with a certain amount of error that was impossible to quantify due to the unsupervised nature of the data (no labels, annotations, or other remarks from any mass production events were available).

Table 4.3: *Kolmogorov-Smirnov* test, *Mann-Whitney U* test and attributes' median values for the production line L4 dataset.

			Median		***Kolmogorov-Smirnov* Test**		***Mann-Whitney U* Test (p-value)**
Cluster Number	**Cluster Size (number and percentage of instances)**	**Attribute Name**	**Cluster**	**NPB**	**Cluster**	**NPB**	
-1	653 1.6%	*Real Gas Consumption*	0.9	10.9	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		CE1 Total Active Energy	12,928.0	16,656.0			
		CE2 Feeder Active Energy	420.0	1,720.0			
		CE1 Reactive Energy	9,032.0	12,255.0			
		CE2 Reactive Energy	288.0	1,242.0			
		CE1 Apparent Energy	9,135.0	11,258.0			
		CE2 Apparent Energy	275.0	1,191.0			
		Entrance Temperature	115.0	200.0			
		Middle Temperature	119.0	200.0			
		Exit Temperature	134.0	199.0			
0	5,634 13.4%	*Real Gas Consumption*	0.0	10.9	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		CE1 Total Active Energy	0.0	16,656.0			
		CE2 Feeder Active Energy	0.0	1,720.0			
		CE1 Reactive Energy	0.0	12,255.0			
		CE2 Reactive Energy	0.0	1,242.0			
		CE1 Apparent Energy	0.0	11,258.0			
		CE2 Apparent Energy	0.0	1,191.0			
		Entrance Temperature	193.0	200.0			
		Middle Temperature	191.0	200.0			
		Exit Temperature	198.0	199.0			
1	28,071 66.7%	*Real Gas Consumption*	-	10.9	-	Non-parametric	-
		CE1 Total Active Energy		16,656.0			
		CE2 Feeder Active Energy		1,720.0			
		CE1 Reactive Energy		12,255.0			
		CE2 Reactive Energy		1,242.0			
		CE1 Apparent Energy		11,258.0			
		CE2 Apparent Energy		1,191.0			
		Entrance Temperature		200.0			
		Middle Temperature		200.0			
		Exit Temperature		199.0			
2	7,523 17.9%	*Real Gas Consumption*	0.0	10.9	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		CE1 Total Active Energy	0.0	16,656.0			
		CE2 Feeder Active Energy	0.0	1,720.0			
		CE1 Reactive Energy	0.0	12,255.0			
		CE2 Reactive Energy	0.0	1,242.0			
		CE1 Apparent Energy	0.0	11,258.0			
		CE2 Apparent Energy	0.0	1,191.0			
		Entrance Temperature	73.0	200.0			
		Middle Temperature	70.0	200.0			
		Exit Temperature	78.0	199.0			
3	136 0.3%	*Real Gas Consumption*	7.5	10.9	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		CE1 Total Active Energy	16,768.0	16,656.0			Statistically equal attributes (0.1)
		CE2 Feeder Active Energy	1,424.0	1,720.0			Statistically different attributes (0.0)
		CE1 Reactive Energy	12,015.0	12,255.0			
		CE2 Reactive Energy	1,148.5	1,242.0			
		CE1 Apparent Energy	11,700.5	11,258.0			
		CE2 Apparent Energy	872.0	1,191.0			
		Entrance Temperature	150.0	200.0			
		Middle Temperature	150.0	200.0			
		Exit Temperature	149.0	199.0			
4	65 0.2%	*Real Gas Consumption*	0.0	10.9	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		CE1 Total Active Energy	14,024.0	16,656.0			
		CE2 Feeder Active Energy	364.0	1,720.0			
		CE1 Reactive Energy	9,241.0	12,255.0			
		CE2 Reactive Energy	70.0	1,242.0			
		CE1 Apparent Energy	10,742.0	11,258.0			
		CE2 Apparent Energy	259.0	1,191.0			
		Entrance Temperature	36.0	200.0			
		Middle Temperature	34.0	200.0			
		Exit Temperature	36.0	199.0			

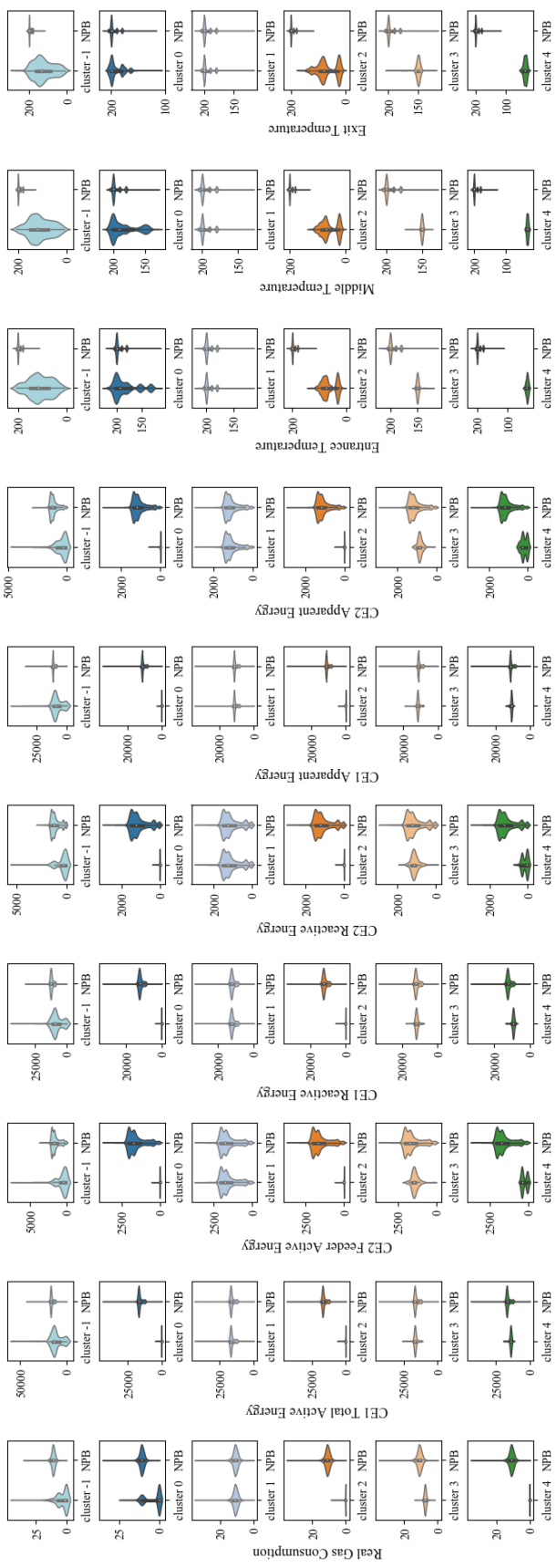

Figure 4.9: Distribution of the process attributes per cluster for the production line L4 dataset.

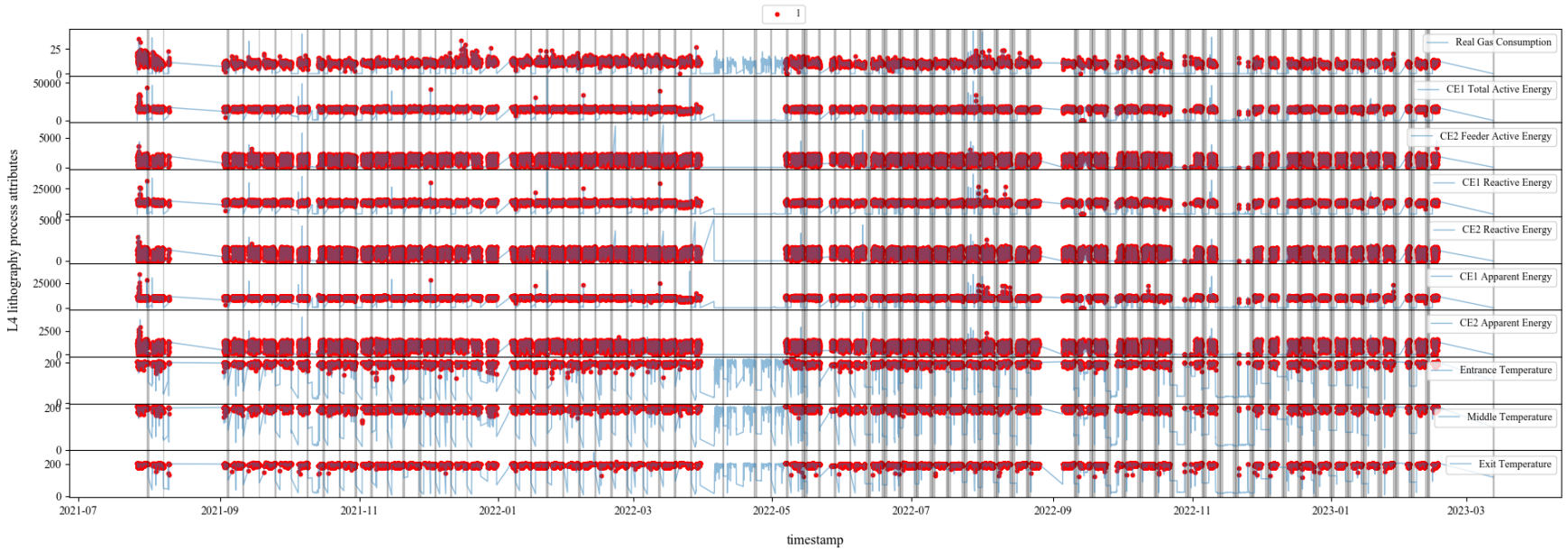

Figure 4.10: Time-series of the process attributes, with the red dots representing instances grouped by cluster 1 (NPB) for the production line L4 dataset.

Table 4.4: *Kolmogorov-Smirnov* test, *Mann-Whitney U* test and attributes' median values for the production line L8 dataset.

			Median		***Kolmogorov-Smirnov* Test**		***Mann-Whitney U* Test (p-value)**
Cluster Number	**Cluster Size (number and percentage of instances)**	**Attribute Name**	**Cluster**	**NPB**	**Cluster**	**NPB**	
-1	1358 4.1%	*Real Gas Consumption*	0.0	10.5	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		Velocity Record	1,715.0	5,010.0			
		Active Energy	8.0	22.0			
		Reactive Energy	6.0	12.0			
		Apparent Energy	10.0	25.1			
		Entrance Temperature	127.0	175.0			
		Middle Temperature 1	161.0	190.0			
		Middle Temperature 2	75.0	80.0			
		Exit Temperature	157.0	190.0			
		Incinerator Temperature	557.0	719.0			
0	453 1.4%	*Real Gas Consumption*	0.0	10.5	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		Velocity Record	10.0	5,010.0			
		Active Energy	0.0	22.0			
		Reactive Energy	0.0	12.0			
		Apparent Energy	0.0	25.1			
		Entrance Temperature	76.0	175.0			
		Middle Temperature 1	84.0	190.0			
		Middle Temperature 2	50.0	80.0			
		Exit Temperature	82.0	190.0			
		Incinerator Temperature	228.0	719.0			
1	139 0.4%	*Real Gas Consumption*	0.0	10.5	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		Velocity Record	13.0	5,010.0			
		Active Energy	0.0	22.0			
		Reactive Energy	0.0	12.0			
		Apparent Energy	0.0	25.1			
		Entrance Temperature	174.0	175.0			
		Middle Temperature 1	191.0	190.0			
		Middle Temperature 2	80.0	80.0			
		Exit Temperature	190.0	190.0			Statistically equal attributes (0.8)
		Incinerator Temperature	563.0	719.0			Statistically different attributes (0.0)
2	1,366 4.2%	*Real Gas Consumption*	7.7	10.5	Non-parametric	Non-parametric	Statistically different attributes (0.0)
		Velocity Record	18.0	5,010.0			
		Active Energy	12.0	22.0			
		Reactive Energy	10.0	12.0			
		Apparent Energy	15.6	25.1			
		Entrance Temperature	175.0	175.0			
		Middle Temperature 1	191.0	190.0			
		Middle Temperature 2	80.0	80.0			
		Exit Temperature	190.0	190.0			
		Incinerator Temperature	561.0	719.0			
3	29,421 89.9%	*Real Gas Consumption*	-	10.5	-	Non-parametric	-
		Velocity Record		5,010.0			
		Active Energy		22.0			
		Reactive Energy		12.0			
		Apparent Energy		25.1			
		Entrance Temperature		175.0			
		Middle Temperature 1		190.0			
		Middle Temperature 2		80.0			
		Exit Temperature		190.0			
		Incinerator Temperature		719.0			

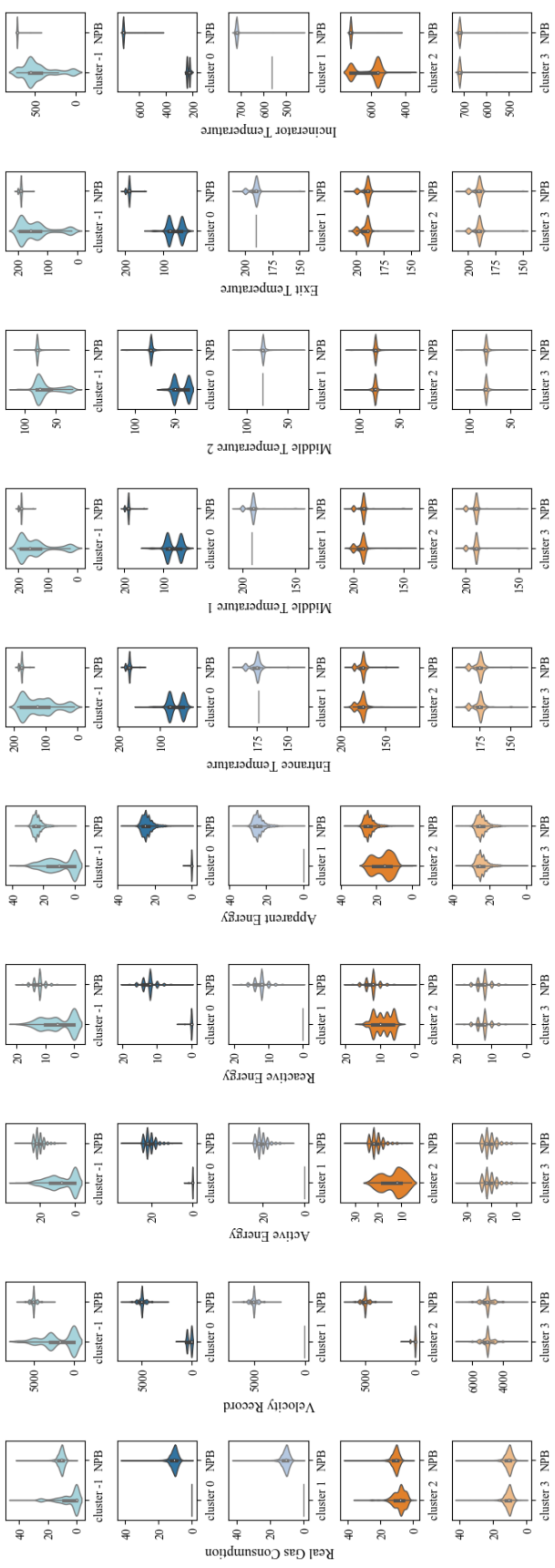

Figure 4.11: Distribution of the process attributes per cluster for the production line L8 dataset.

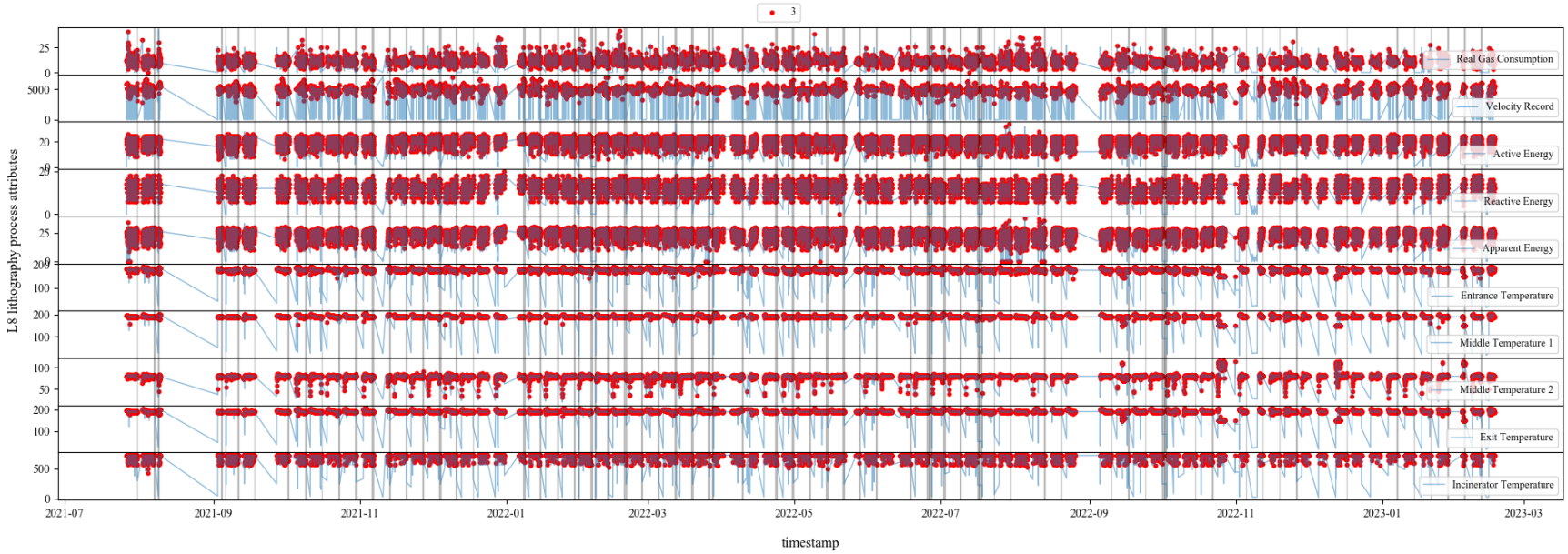

Figure 4.12: Time-series of the process attributes, with the red dots representing instances grouped by cluster 3 (NPB) for the production line L8 dataset.

Table 4.5: Train and test split schema for the production line L4 and L8 datasets, with the percentage of the total number of instances between brackets.

	Production Line	
	L4	**L8**
Total number of instances	**42,082 (100.0%)**	**32,737 (100.0%)**
Total NPB instances	28,071 (66.7%)	29,572 (90.3%)
NPB instances for training	19,649 (70.0%)	20,700 (70.0%)
NPB instances for testing	8,422 (30.0%)	8,872 (30.0%)
Total APB instances	14,011 (33.3%)	3,165 (9.7%)
Total number of instances use for training	19,649 (46.7%)	20,700 (63.2%)
Total number of instances use for testing	22,433 (53.3%)	12,037 (36.8%)

4.4 Outlier Detection

Finished the process mapping, the NPB cluster was clearly identified for the production lines L4 and L8 datasets. The application of OD algorithms capable of learning what is established as the NPB is now possible. In this way, OD algorithms could be used to continuously monitor the lithography process, searching for outliers that might be related to process anomalies.

Note that OD algorithms could technically be used after PCA application. Nevertheless, they will only be able to split the dataset's instances between two different clusters (inliers or outliers), making the task of identifying the different process operating modes extremely difficult.

The production line L4 and L8 datasets were split according to the train and test schema presented in Table 4.5. The applied OD algorithms should only be trained with instances from the NPB cluster. Instances from the APB clusters will only be used during testing. Although OD algorithms are applied to what is considered a normal process, we are aware that the obtained NPB clusters might have included some amount of outliers. The contamination hyperparameter of each OD model will handle this technical obstacle.

Three different OD algorithms were used to learn the NPB frontier, namely, Isolation Forest (IF), Local Outlier Factor (LOF) and One-Class Support Vector Machine (OCSVM). A grid-search approach through the selected hyperparameters was adopted to find the best model for each production line dataset. The models were evaluated using two different metrics, F_1 Score and ROC AUC. A visual self-evaluation of the learned frontier was also performed, divided into two categories (overfitted or underfitted model). The results obtained for the best models are presented in Table 4.6. A mean and standard deviation value is shown due to the random seed given to the method (10 random seeds were used for each OD model).

Figures from Figure 4.13 to Figure 4.24 present the learned frontiers (dark red line) of each best model from Table 4.6. True inliers (or true positives) and true outliers (or true negatives) are also identified in the figures. The true inlier and outlier nomenclature were chosen since we considered them more adequate to the nature of our problem. The true inliers (dark blue points) represent instances from the NPB cluster, while true outliers (light blue points) represent instances from the APB cluster. False positives (light red points) and false negatives (light orange points) will also influence the model's performance. Therefore, they were also represented in the figures.

Table 4.6: Evaluation metrics of the best OD models for production line L4 and L8 datasets.

	Evaluation Metrics (Visual Self-Evaluation)			
	Production Line L4		Production Line L8	
Outlier Detection Algorithm	**F_1 Score**	**ROC AUC**	**F_1 Score**	**ROC AUC**
Isolation Forest	0.9964 ± 0.0003 (u)	0.9968 ± 0.0004 (u)	0.9918 ± 0.0008 (u)	0.9877 ± 0.0010 (u)
Local Outlier Factor	0.9974 ± 0.0005 (o)	0.9979 ± 0.0004 (o)	0.9879 ± 0.0007 (o)	0.9731 ± 0.0011 (o)
One-Class Support Vector Machine	**0.9996 ± 0.0001 (s)**	**0.9997 ± 0.0001 (s)**	**0.9983 ± 0.0004 (s)**	**0.9980 ± 0.0003 (s)**

(o) - Visual overfitted model
(s) - Selected model
(u) - Visual underfitted model

For clarification, false positives are instances from the APB cluster (outliers) that the model wrongly classified as being part of the NPB cluster (inliers). False negatives are instances from the NPB cluster (inliers) that the model wrongly classified as being part of the APB cluster (outliers).

Overall, all the best models obtained perform considerably well for the two evaluation metrics used. A detailed evaluation of each model will be promptly fulfilled.

Concerning the production line L4 dataset, the OCSVM models presented the best performance at both evaluation metrics (0.9996 and 0.9997 for the F_1 Score and ROC AUC, respectively), together with the frontier's visual self-evaluation (Figure 4.17 and Figure 4.18). Although the LOF models have also presented considerable good results (0.9974 and 0.9979 for the F_1 Score and ROC AUC, respectively), after a visual self-evaluation, the learned frontier presented some minor overfitted areas, particularly on the top area of the NPB cluster (Figure 4.15 and Figure 4.16). The IF models presented the lowest performance (0.9964 and 0.9968 for the F_1 Score and ROC AUC, respectively), with the best-obtained models presenting an underfitted area on the bottom of the NPB cluster (Figure 4.13 and Figure 4.14).

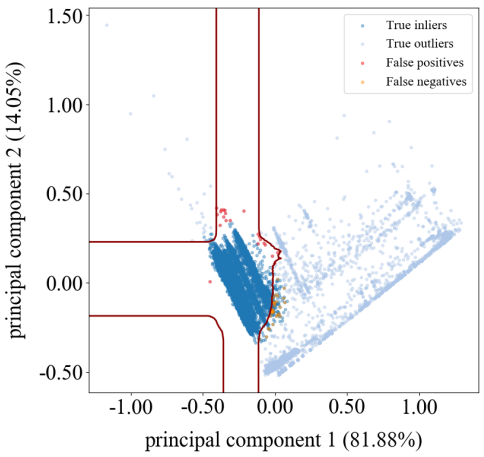

Figure 4.13: Best F_1 Score (0.9964) IF model for the production line L4 dataset. The model was obtained with a *number of estimators* of 800 and a *contamination* of 0.0053.

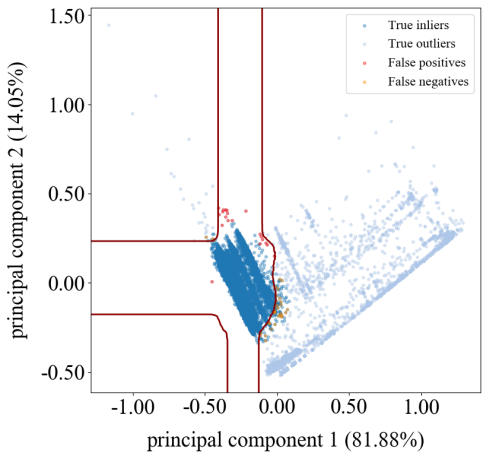

Figure 4.14: Best ROC AUC (0.9968) IF model for the production line L4 dataset. The model was obtained with a *number of estimators* of 800 and a *contamination* of 0.0051.

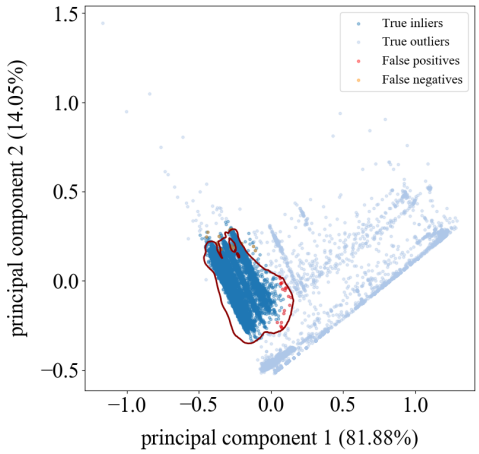

Figure 4.15: Best F_1 Score (0.9974) LOF model for the production line L4 dataset. The model was obtained with a *number of neighbours* of 104, *contamination* of 0.0036, and considering the *novelty detection* approach.

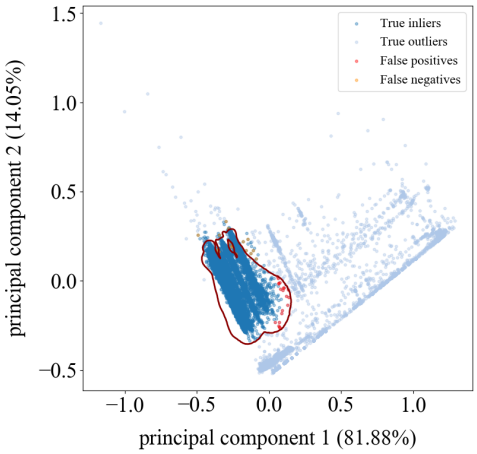

Figure 4.16: Best ROC AUC (0.9979) LOF model for the production line L4 dataset. The model was obtained with a *number of neighbours* of 117, *contamination* of 0.0031, and considering the *novelty detection* approach.

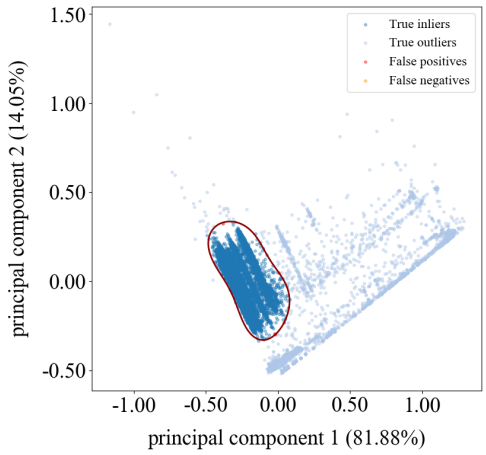

Figure 4.17: Best F_1 Score (0.9996) OCSVM model for the production line L4 dataset. The model was obtained with a *nu* of 0.0004.

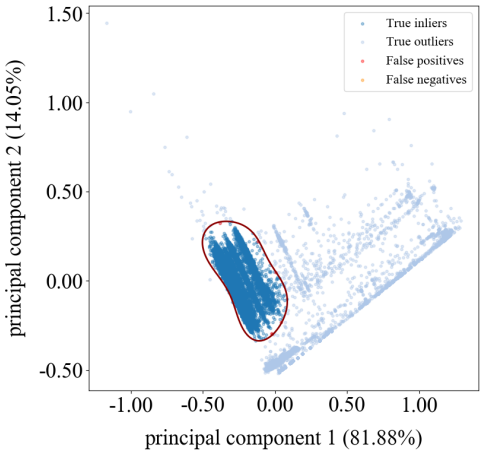

Figure 4.18: Best ROC AUC (0.9997) OCSVM model for the production line L4 dataset. The model was obtained with a *nu* of 1.3×10^{-5}.

Concerning the production line L8 dataset, the OCSVM models presented the best and most similar performance at both evaluation metrics (0.9983 and 0.9980 for the F_1 Score and ROC AUC, respectively), together with the frontier's visual self-evaluation (Figure 4.23 and Figure 4.24). Although the IF models have presented considerable good results in the testing phase (0.9918 and 0.9877 for the F_1 Score and ROC AUC, respectively), the learned frontier is not particularly good in isolating the NPB cluster (Figure 4.19 and Figure 4.20). This is due to open areas in the frontier (parallel dark red lines), increasing the model's false positive detection. We are convinced that a decrease in the IF models' performance would be noticed if additional APB instances were

available in the testing phase. The LOF models presented the lowest performance (0.9879 and 0.9731 for the F_1 Score and ROC AUC, respectively), with the best-obtained models presenting some overfitted areas along the learned frontier (Figure 4.21 and 4.22).

Note that, visually, both OCSVM and LOF models have similar frontier shapes. Nevertheless, the ones obtained through the OCSVM model present a smoother, thus less overfitted contour. It is also noted that the learned frontiers showed different shapes for the same model. This is because they are represented in different principal component projections.

In Figure 4.19 to Figure 4.22, it is possible to identify some false positives (light red points) in dense areas of the NPB cluster. This presents a limitation of the cluster model chosen since, in some regions, an overlap between clusters can occur.

Although the obtained models presented particularly good results (close to 1), some limitations should be taken into consideration. Firstly, no tracking of the NPB instances used for testing was performed, which means that these instances were randomly chosen, not being possible to guarantee their location in the feature space. Instances located close to the learned frontier will certainly be more difficult to classify, and therefore a decrease in the model's overall performance will be noticeable. Secondly, based on the nature of our problem, there is a decrease in the instances' density close to the learned frontier. Regions with a low-density instance will lead to fewer instances to classify, decreasing the probability of misclassification. Lastly, for the production line L8, an additional effort should be performed to obtain additional APB instances to better balance both categories (APB and NPB) in the dataset.

Figure 4.25 and Figure 4.27 present instances grouped in two categories, inliers and outliers, based on the classification performed by the best OCSVM model (ROC AUC metric). The anomaly score was also added to the time-series, which quantifies the degree of anomaly of an instance. Instances with a higher anomaly score than the threshold are considered outliers (red points). Conversely, instances with a lower anomaly score than the threshold are considered inliers (green points). In future developments, this parameter could support the control operator during the fine-tuning of the production equipment.

In Figure 4.26 and Figure 4.28, a detailed visualization from the previous Figure 4.25 and Figure 4.27 is presented concerning the period between 2022-09-05 to 2022-09-28 (colour-coded in light yellow). In both figures, regions where the lithography process was stable can be identified, with the instances being correctly classified as inliers by the model. The model is also capable of highlighting global outliers (Figure 4.26 on period 2022-09-23, window A) and collective outliers (Figure 4.28 on period 2022-09-22, window B). For the production line L4, it is also clear that the model can identify the equipment shutdown periods during the weekends (colour-coded in light grey).

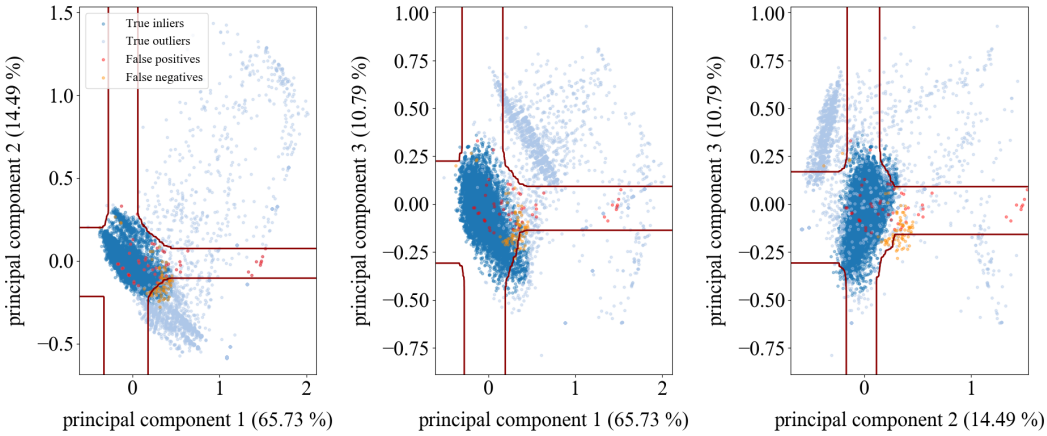

Figure 4.19: Best F_1 Score (0.9918) IF model for the production line L8 dataset. The model was obtained with a *number of estimators* of 200 and a *contamination* of 0.0106.

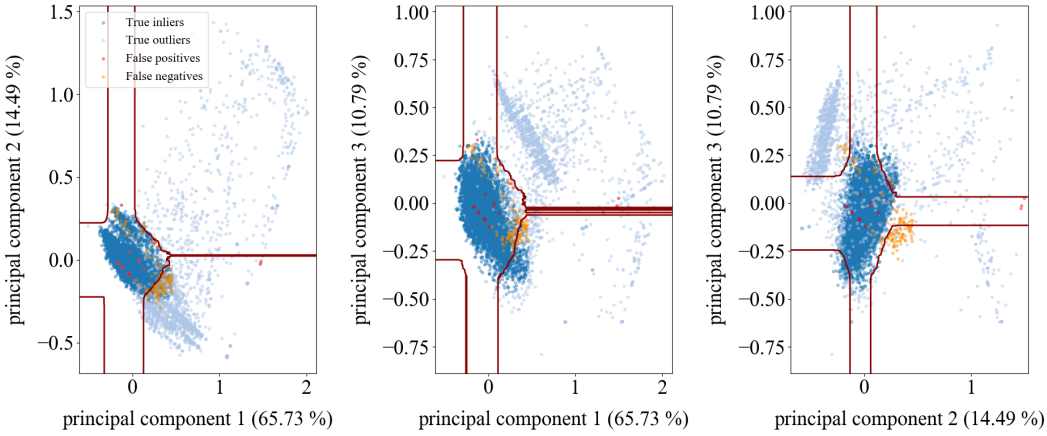

Figure 4.20: Best ROC AUC (0.9877) IF model for the production line L8 dataset. The model was obtained with a *number of estimators* of 200 and a *contamination* of 0.0157.

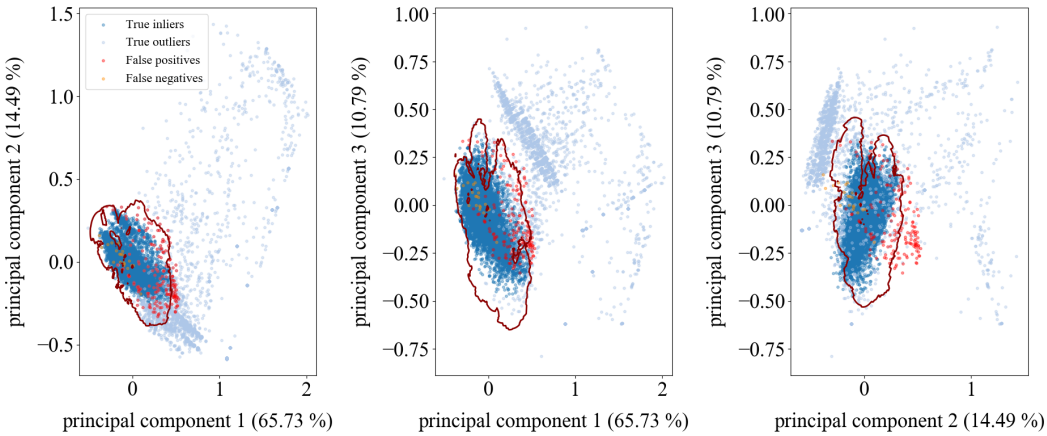

Figure 4.21: Best F_1 Score (0.9879) LOF model for the production line L8 dataset. The model was obtained with a *number of neighbours* of 5, *contamination* of 0.0042, and the *novelty detection* approach.

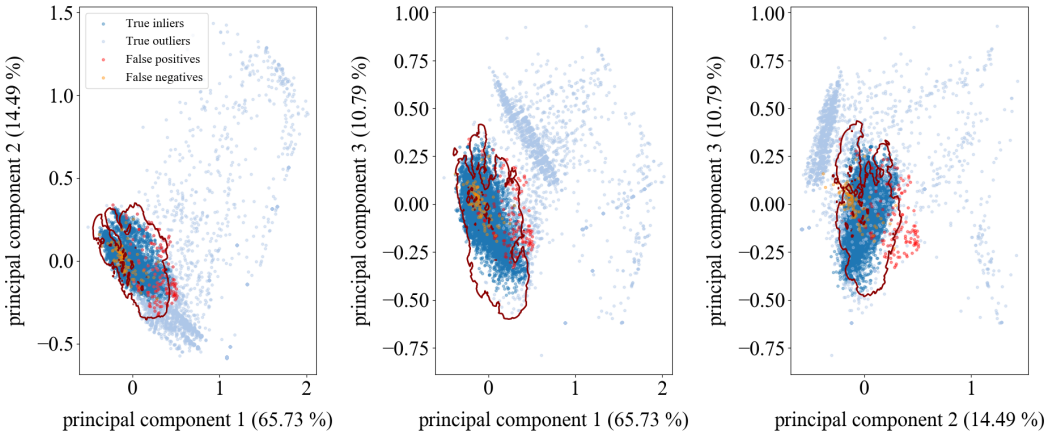

Figure 4.22: Best ROC AUC (0.9731) LOF model for the production line L8 dataset. The model was obtained with a *number of neighbours* of 4, *contamination* of 0.0118, and the *novelty detection* approach.

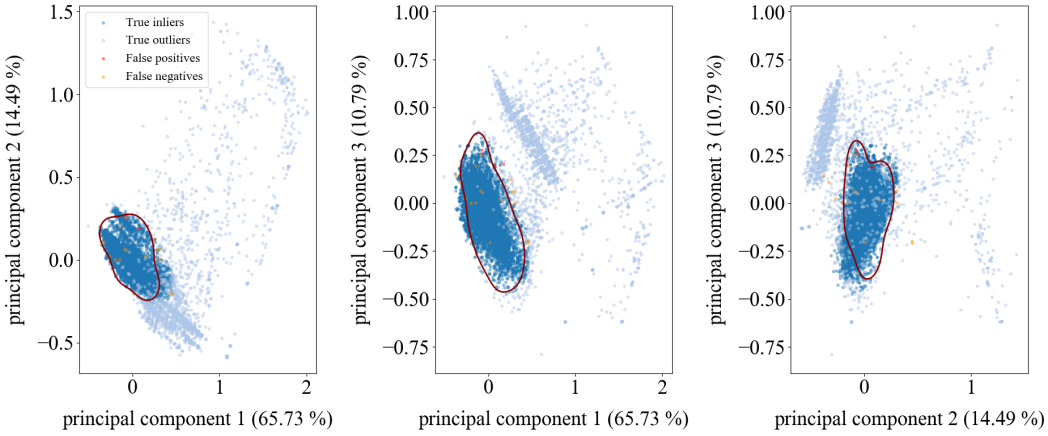

Figure 4.23: Best F_1 Score (0.9983) OCSVM model for the production line L8 dataset. The model was obtained with a *nu* of 0.0008.

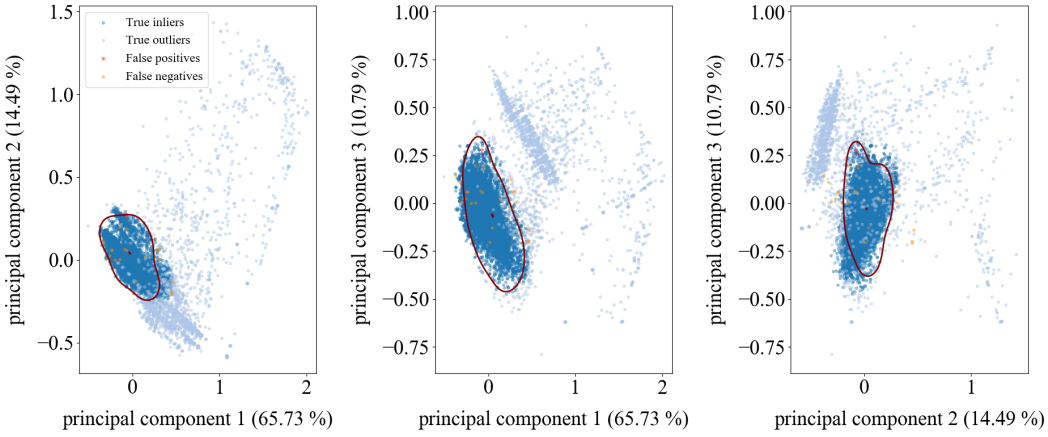

Figure 4.24: Best ROC AUC (0.9980) OCSVM model for the production line L8 dataset. The model was obtained with a *nu* of 0.0022.

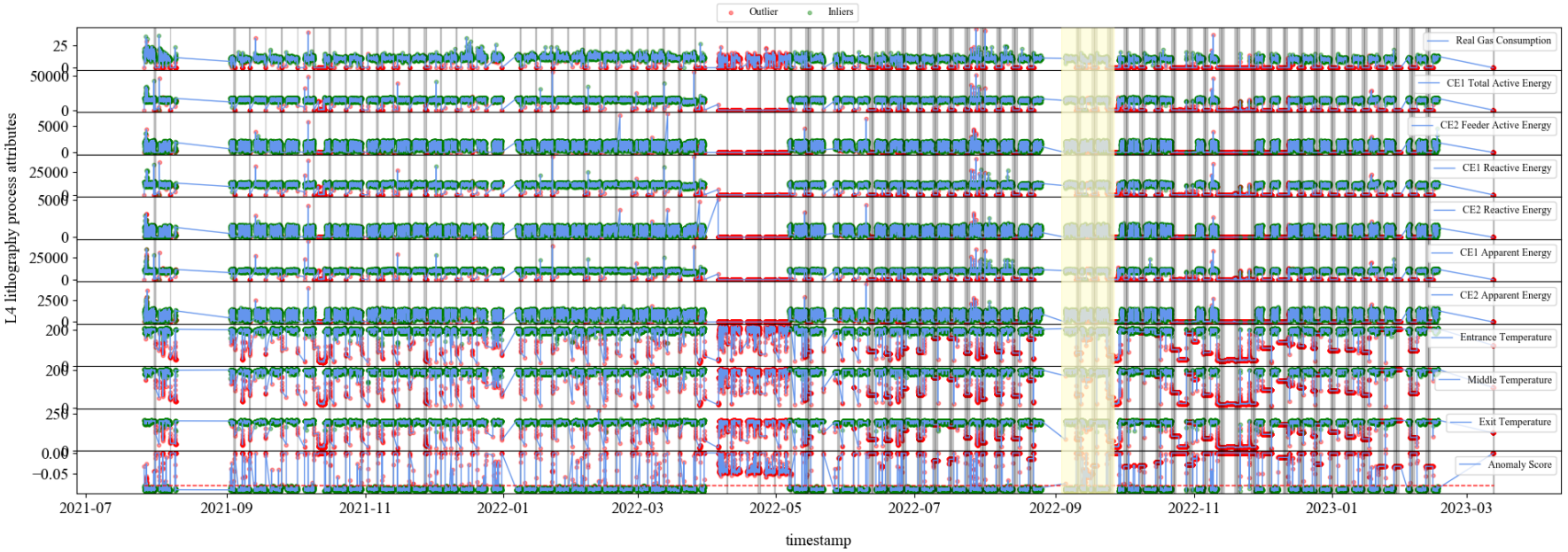

Figure 4.25: Time-series with the inliers (green dots) and outliers (red dots) classified by the best OCSVM model for the production line L4 dataset.

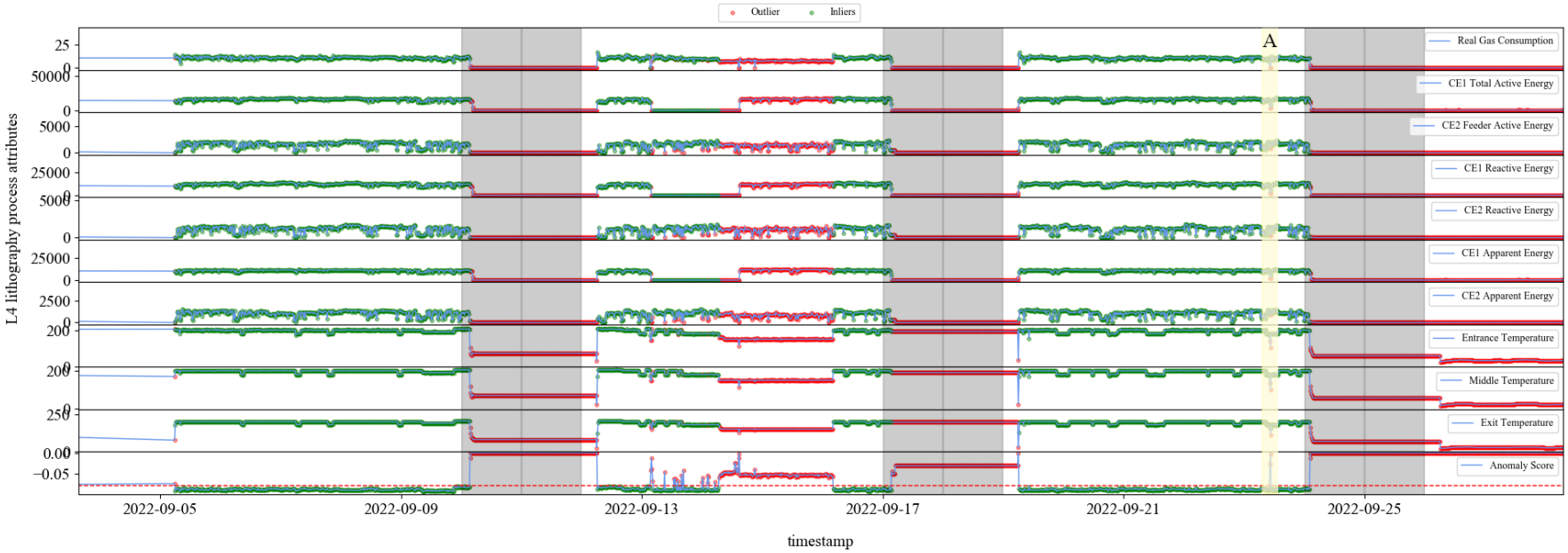

Figure 4.26: A detail from the time-series presented in Figure 4.25.

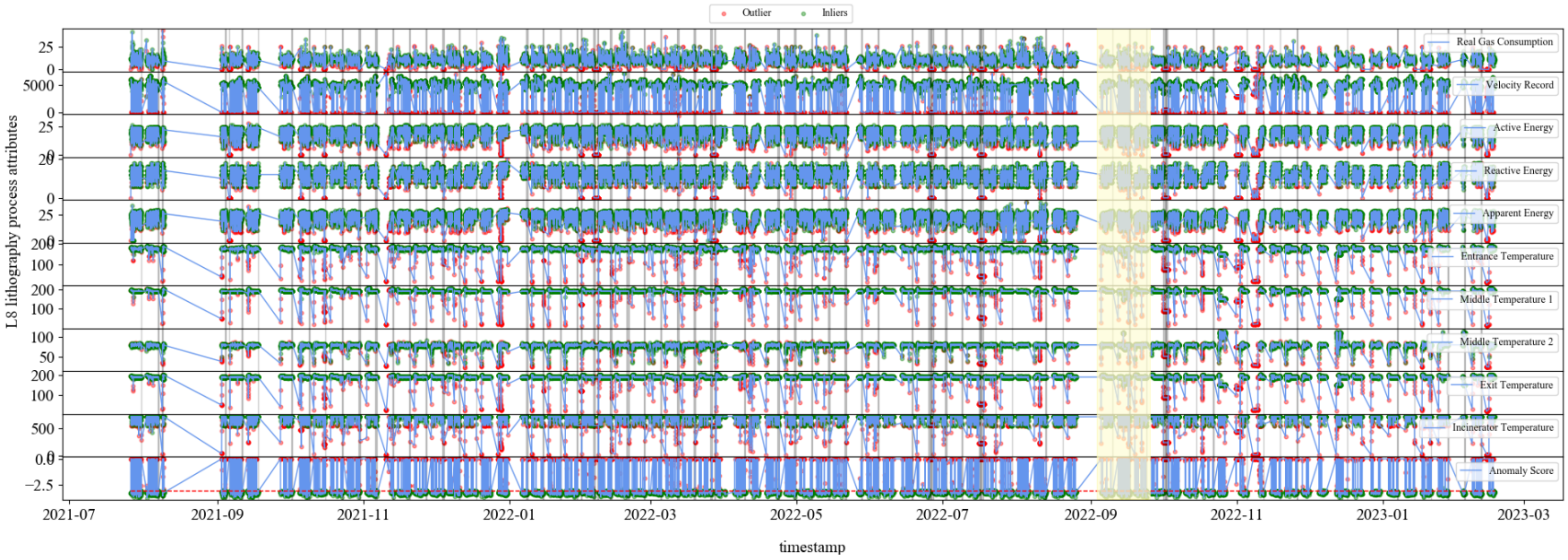

Figure 4.27: Time-series with the inliers (green dots) and outliers (red dots) classified by the best OCSVM model for the production line L8 dataset.

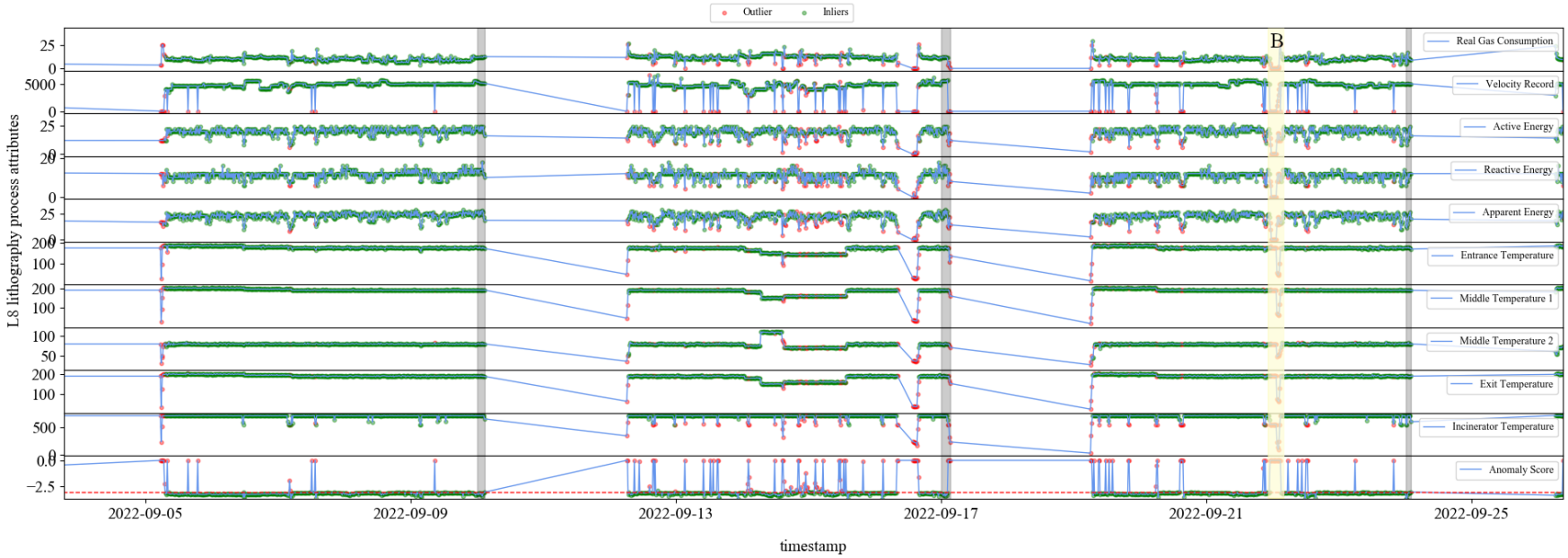

Figure 4.28: A detail from the time-series presented in Figure 4.27.

Chapter 5

Conclusions and Future Work

During this work, unlabeled multivariate time-series from two different lithography production lines (L4 and L8) were studied. The best principles to maximise the datasets' quality without jeopardising their size were adopted during the data processing. A PCA was performed as a dimensionality reduction technique to decrease the problem's complexity and allow a visual self-evaluation of the future clustering models. Ultimately, the L4 and L8 production lines datasets were split into two and three principal components, respectively, with a variance explainability above 90%.

Different clustering algorithms were applied to identify clusters in the dimensional-reduced datasets. For the production line L4 and L8 datasets, the best results were achieved when DBSCAN and HDBSCAN were applied, respectively. The obtained clusters were then fully mapped to identify the different lithography process operating modes, and the NPB cluster was identified. With the application of OD algorithms, the NPB was learned, and the best results were achieved for OCSVM-based models.

Based on the previously stated and concerning the research questions proposed, anomalous lithography production patterns can be detected using PCA together with the application of OD algorithms. Regarding the performance of the OD algorithms, for the production line L4 dataset, an F_1 Score and ROC AUC of 99.96% and 99.97% were achieved, respectively. Concerning the production line L8 dataset, an F_1 Score and ROC AUC of 99.83% and 99.80% were achieved, respectively.

As further work, we suggest the exploration of four main topics, namely:

- **Relationship between detected outliers and non-conforming lithography batches**: a close interaction with Colep is recommended since additional quality history should be provided to crosscheck the model's detected outliers with the actual non-conforming batches;

- **Feature engineering**: the supervision of a process engineer with specific lithography knowledge is recommended to bring additional process attributes to the available dataset;

- **Feature importance**: to obtain additional insights from the model's outputs, local and global explainability should be further explored. A model-agnostic approach based on Shapley values might address this topic (15). Since a PCA was initially adopted, the obtained explainability will be related to the selected principal components, which would not have a straightforward interpretation. Therefore, a different data modelling approach might be needed to bring the model's explainability to the shop floor level;

- **Pipeline implementation**: this work is able to support the development of a pipeline for a real-scenario application. Firstly, new incoming data must be projected according to the principal components of interest. Secondly, the selected OCSVM model must fit the new projected data. In the end, the model will provide an inlier or outlier classification.

Based on the developed work, an anomalous pattern detection methodology applied to a lithography process was achieved. This approach supported the metallic packing industry's entry into Industry 4.0 by using ML techniques for continuous process monitoring tasks. In the future, we expect to prevent non-conformities and support quality control during mass production events.

References

[1] AGGARWAL, C. C. *Outlier Analysis*, 2 ed. Springer Cham, 2016.

[2] AMINIKHANGHAHI, S., AND COOK, D. J. A survey of methods for time series change point detection. *Knowledge and Information Systems 51* (2017), 339–367.

[3] BREUNIG, M. M., KRIEGEL, H.-P., NG, R. T., AND SANDER, J. Lof: Identifying density-based local outliers. pp. 93–104.

[4] CHANDOLA, V., BANERJEE, A., AND KUMAR, V. Anomaly detection: A survey. *Association for Computing Machinery 41* (2009).

[5] DAVIES, D. L., AND BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1*, 2 (1979), 224–227.

[6] DESGRAUPES, B. Clustering indices. University Paris Ouest Lab Modal'X.

[7] GOLDSTEIN, M., AND DENGEL, A. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. pp. 59–63.

[8] HE, Z., XU, X., AND DENG, S. Discovering cluster-based local outliers. *Pattern Recognition Letters 24* (2003), 1641–1650.

[9] JAVAID, M., HALEEM, A., SINGH, R. P., AND SUMAN, R. Significance of quality 4.0 towards comprehensive enhancement in manufacturing sector. *Sensors International 2* (2021).

[10] LAI, K.-H., ZHA, D., XU, J., ZHAO, Y., WANG, G., AND HU, X. Revisiting time series outlier detection: Definitions and benchmarks.

[11] LASHKOV, A., AND GUPTA, S. S_dbw validity index. https://github.com/alashkov83/S_Dbw.

[12] LIU, F. T., TING, K. M., AND ZHOU, Z. H. Isolation forest. pp. 413–422.

[13] LIU, Y., LI, Z., XIONG, H., GAO, X., AND WU, J. Understanding of internal clustering validation measures. pp. 911–916.

[14] MCINNES, L., HEALY, J., AND ASTELS, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software 2*, 11 (2017), 205.

[15] MOLNAR, C. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*, 2 ed. 2022.

[16] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[17] PIMENTEL, M. A., CLIFTON, D. A., CLIFTON, L., AND TARASSENKO, L. A review of novelty detection. *Signal Processing 99* (2014), 215–249.

[18] PINHO, P. Criação e sistematização de um pco (product cost optimization) na litografia, 2014.

[19] RAMASWAMY, S., RASTOGI, R., AND SHIM, K. Efficient algorithms for mining outliers from large data sets. pp. 427–438.

[20] REIS, M. *Estatística para a Melhoria de Processos: A Perspectiva Seis Sigma*, 1 ed. 2016.

[21] ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20* (1987), 53–65.

[22] SCHMIDL, S., WENIG, P., AND PAPENBROCK, T. Anomaly detection in time series: A comprehensive evaluation. vol. 15, VLDB Endowment, pp. 1779–1797.

[23] SCHÖLKOPF, B., WILLIAMSON, R. C., SMOLA, A., SHAWE-TAYLOR, J., AND PLATT, J. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems* (1999), S. Solla, T. Leen, and K. Müller, Eds., vol. 12, MIT Press, pp. 582–588.

[24] SCIKIT-LEARN. Clustering. https://scikit-learn.org/stable/modules/clustering.html.

[25] SCIKIT-LEARN. DBSCAN. http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html.

[26] SCIKIT-LEARN. Gaussian Mixture. https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html.

[27] SCIKIT-LEARN. Isolation Forest. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html.

[28] SCIKIT-LEARN. Local Outlier Factor. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html.

[29] SCIKIT-LEARN. Mean Shift. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html.

[30] SCIKIT-LEARN. One Class SVM. https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html.

[31] TAX, D. *One-class classification: Concept-learning in the absence of counter-examples*. PhD thesis, Technische Universiteit Delft, 2001.

[32] VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., VAN DER WALT, S. J., BRETT, M., WILSON, J., MILLMAN, K. J., MAYOROV, N., NELSON, A. R. J., JONES, E., KERN, R., LARSON, E., CAREY, C. J., POLAT, İ., FENG, Y., MOORE,

E. W., VANDERPLAS, J., LAXALDE, D., PERKTOLD, J., CIMRMAN, R., HENRIKSEN, I., QUINTERO, E. A., HARRIS, C. R., ARCHIBALD, A. M., RIBEIRO, A. H., PEDREGOSA, F., VAN MULBREGT, P., AND SCIPY 1.0 CONTRIBUTORS. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17* (2020), 261–272.

[33] WANG, B., AND MAO, Z. Outlier detection based on gaussian process with application to industrial processes. *Applied Soft Computing Journal 76* (2019), 505–516.

[34] ZHAO, Y., NASRULLAH, Z., AND LI, Z. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research 20* (2019), 1–7.