A Novel Multi-perspective Trace Clustering Technique for IoT-Enhanced Processes: A Case Study in Smart Manufacturing

Yannis Bertrand(✉), Jochen De Weerdt, and Estefanía Serral

Research Centre for Information Systems Engineering (LIRIS), KU Leuven, Warmoesberg 26, 1000 Brussels, Belgium
{yannis.bertrand,jochen.deweerdt,estefania.serralasensio}@kuleuven.be

Abstract. *IoT-enhanced business processes (BPs)* are processes supported by Internet of Things (IoT) technology, such as sensors capable of monitoring the physical environment where processes are executed. Although the execution of BPs is typically recorded in event logs, IoT-enhanced BPs also generate IoT data that contain vital contextual information. Such BPs are typically found in manufacturing contexts, where, for instance, temperature sensors can provide valuable insights into the storage conditions of sensitive raw materials. However, the potential of this *IoT-enhanced process mining (PM)* has not been fully explored. In this paper, we propose TROPIC, an approach for multi-perspective trace clustering that considers three key perspectives: the control-flow perspective, the trace attribute data perspective and the time series sensor data perspective. We demonstrate the efficacy of our approach in a real-world manufacturing use case. The evaluation of the resulting clusters revealed that integrating the three different perspectives enabled the detection of process variants and anomalous instances that would have been missed using any one of the perspectives in isolation.

Keywords: Process mining · Internet of Things · Trace clustering · IoT-enhanced process mining

1 Introduction

Currently, the use of Internet of Things (IoT) devices in organisations is becoming increasingly common, providing support to their business processes (BPs), known as *IoT-enhanced BPs* [16,36]. The execution of BP activities is usually recorded in event logs, which can be analysed to gain insights into the BP and identify opportunities for improvement. When BPs are augmented with IoT devices, these devices can also provide critical contextual information. One of the main domains where IoT-enhanced BPs are found is smart manufacturing. In these BPs, sensors can track time series (TS) data on various process parameters, such as, for example, flow, temperature, and pressure, which can aid in

C. Di Francescomarino et al. (Eds.): BPM 2023, LNCS 14159, pp. 395–412, 2023.
https://doi.org/10.1007/978-3-031-41620-0_23

predicting process outcomes and automating tasks. However, due to the unique characteristics of IoT data, such as granularity and storage independent of the process system [4], it is necessary to develop new PM techniques designed specifically for them. This emerging field of *IoT-enhanced process mining (PM)* is still in its early stages [4], with only limited research being already done, focusing primarily on decision mining using IoT data [2,32].

In this paper, we propose TROPIC (TRace attributes, cOntrol-flow Plus Iot Clustering), a novel approach for multiperspective trace clustering that is capable of integrating the TS sensor data perspective, in addition to the control-flow and trace attribute data perspectives. By integrating these different perspectives, multi-perspective trace clustering can effectively identify process variants and anomalous process executions in smart manufacturing that may not be apparent from analysing the control-flow or another single perspective alone. Knowing these variants can, in turn, help organisations identify and propagate best practises to enhance process efficiency and increase the likelihood of positive process outcomes. To demonstrate the effectiveness of our approach, we apply it to a real-life manufacturing process and provide a detailed evaluation of the results. This case study highlights the potential of our approach to analyse and improve IoT-enhanced BPs.

The remainder of the paper is organised as follows. First, Sect. 2 provides an overview of previous research in multi-perspective PM, IoT-enhanced PM, and trace clustering. In Sect. 3, we present TROPIC, our two-level approach for multi-perspective trace clustering, and apply it to the manufacturing process in question in Sect. 4. The experimental results are discussed in Sect. 5, before concluding the paper in Sect. 6 with final remarks and suggestions for future work.

2 Background

2.1 Trace Clustering

Trace clustering is a technique used to group similar process instances, for instance, based on their shared sequential activity patterns. Traditionally, trace clustering has been used to improve process discovery by splitting the event log into sublogs consisting of instances that share comparable activity sequences, and mining a model of each sublog separately. This approach produces simpler and better fitting models that describe different process variants [5,9,13]. However, more recently, trace clustering has been applied to other goals, such as concept drift detection and process evolution analysis [19] and outlier detection [11]. Although improving process discovery results can typically rely only on the control-flow perspective, other objectives can greatly benefit from incorporating context information in clustering.

According to [8], three main categories of trace clustering approaches have been proposed: distance-based, feature-based, and model-based. Distance-based approaches directly cluster traces based on the distances between traces as sequences of activities, using distance metrics such as the Hamming distance,

Levenshtein distance, Damerau-Levenshtein distance, and geodesic distance. Feature-based techniques, on the other hand, derive features from the traces, such as scalars, graphs and embeddings and cluster based on the feature values. Finally, model-based techniques aim to create clusters of traces that produce the best process models [9], optimising criteria such as model fitness. These three approaches have their advantages and disadvantages depending on the nature of the data and the intended application. Choosing the appropriate approach is critical to the effectiveness of the trace clustering process.

2.2 Multi-perspective Process Mining

Multi-perspective PM refers to process analysis techniques that take more than one process perspective into account, e.g., the control-flow and data attributes. The following perspectives are listed in [22] lists the following perspectives:

- Control-flow perspective: Activity ordering in each process instance;
- Resource perspective: Human and non-human resources executing tasks;
- Data perspective: Trace and event attributes;
- Time perspective: Activity duration, throughput time, business rules, etc.;
- Function: Granularity of the activities of the process.

Multi-perspective techniques have been proposed for various types of PMs, such as multi-perspective process discovery [18,24] and multi-perspective conformance checking [14,23]. In trace clustering, a multi-perspective approach is proposed in [15], where a distance metric is presented to compare traces based on the control-flow perspective, the resource perspective, and the data perspective. The (possibly weighted) average of these metrics is computed and used as a pairwise multi-perspective distance measure to perform hierarchical clustering.

However, extending such a technique to TS data can be challenging, as TS typically need to be characterised by many features. For example, [12] reviewed the proposed TS characteristics in the literature and identified a list of approximately 7,700 characteristics to fully represent the TS data. Therefore, proceeding in one step, inputting TS features in a feature vector or including them in an average as in [15], would likely result in either TS features dominating over other perspectives or require very carefully selecting TS features beforehand. This problem grows dramatically when considering multivariate TS, which are very common in manufacturing. To address this issue, we propose a two-step approach that is more versatile than the simple average of distances computed over multiple perspectives.

2.3 IoT-Enhanced PM

Event Log Derivation. The existing literature on IoT-enhanced PM has primarily focused on deriving high-level events of the process from low-level IoT data to create event logs. Subsequently, traditional PM techniques have been employed to analyse these event logs and discover control-flow models of the

processes. Several techniques have been proposed specifically for manufacturing processes. In [35], a four-step framework is presented to generate event logs from industrial IoT data, including data preprocessing, clustering of low-level data, classification to derive events from clusters, and creation of the final event log. Also, focussing on industrial applications, [34] propose to transform raw IoT data into an XES event log using complex event processing and event detection and refinement techniques. The authors present another approach in [33] to detect activities interactively from sensor data based on visualisation and exploratory analysis. In [37], a domain-specific language is developed to extract event logs from IoT data by specifying the case and activity identifiers.

Process Contextualisation. Next to event log derivation, some context-aware techniques have also been investigated, e.g., IoT data-aware process discovery [2,20], sensor TS-aware decision mining [32], and IoT-aware conformance checking [28]. In a manufacturing context, [32] outlines an approach to derive decision rule patterns from TS sensor data by automatically featurising the sensor data and training a decision tree to learn the rules. A different problem is addressed by [28], who present an approach for IoT-enhanced deviation detection. In their paper, they argue that traditional conformance checking cannot take into account data that changes over time independently of the events of the process (i.e., TS data). They subsequently proposed a method to detect patterns in the TS data directly.

3 Methodology

TROPIC involves a two-step clustering process (see Fig. 1) currently tailored to the setting of smart manufacturing, typically characterised by highly structured processes around which sensor data are collected in the form of TS. Indeed, in such manufacturing BPs, sensor data and process activities are usually correlated, with process activities leaving recognisable patterns in the sensor data and certain sensor data values triggering the execution of certain process activities. In the clustering process of TROPIC, process instances are first clustered *separately* according to three perspectives: the *control-flow*, *trace attribute data* and *TS sensor data* perspectives. In this step, each perspective is considered independently, providing a detailed view of each aspect of the process. Then, the distances to each centroid in each clustering are computed and used as features for a second clustering step, which takes into account all three perspectives together. This results in a multi-perspective clustering that groups instances based on their unique combinations of control-flow, trace attributes and TS sensor data, providing a comprehensive view on the process.

Next, we explain the approach applied for each single-perspective clustering, followed by the multi-perspective clustering.

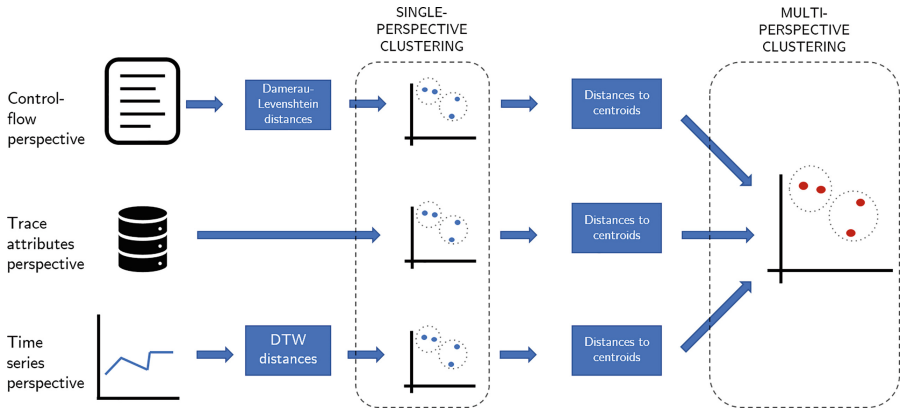


Fig. 1. Overview of TROPIC, our proposed approach.

3.1 Control-Flow Perspective

As mentioned in Sect. 2, three main categories of trace clustering have been proposed: distance-based, feature-based, and model-based approaches. Our approach follows the former by using the Damerau-Levenshtein (DL) distance. The DL distance is a string metric used to compute the edit distance between two strings, which is the minimum number of single-character edits (i.e., insertions, deletions, substitutions, and transpositions) required to transform one string into the other. It extends the Levenshtein distance by also including transpositions of characters. The DL distance between strings A and B, denoted DL(A,B), is computed as follows:

$$DL(A,B) = \begin{cases} \max(|A|,|B|) & \text{if } \min(|A|,|B|) = 0 \\ \min \begin{cases} DL(A_{1..i-1},B)+1 \\ DL(A,B_{1..j-1})+1 \\ DL(A_{1..i-1},B_{1..j-1})+\delta_{a_i,b_j} \\ DL(A_{1..i-2}b_i,A_{1..j-2}a_j)+1 \end{cases} & \text{otherwise} \end{cases} \quad (1)$$

where $|A|$ denotes the length of string A, a_i denotes the i-th character of string A, and δ_{a_i,b_j} is the Kronecker delta function, which is equal to 1 if $a_i = b_j$, and 0 otherwise. The last term in the minimum function corresponds to transposition, and is only included if $i, j > 1$ and $a_{i-1} = b_j$ and $b_{j-1} = a_i$.

Due to the strictly ordered nature of control-flow data in many manufacturing processes, other trace clustering paradigms are usually less suitable. Additionally, activities are often logged at a fairly low level of granularity, making model-based techniques less appropriate. It is worth noting that manufacturing processes tend to be more structured in nature, and thus may not require more complex trace clustering techniques designed for less structured processes.

3.2 Trace Attribute Data Perspective

Trace attributes are usually numerical, categorical, or ordinal features that can be clustered using traditional clustering techniques. Common clustering techniques include hierarchical techniques [38], distance-based techniques, such as K-means [21] or K-medoids [27], model-based techniques, such as self-organising maps [17]; and density-based techniques such as DBSCAN [10]. TROPIC uses K-means, as a generic technique for mixed-type input features, which is most often the case in smart manufacturing. Moreover, its simplicity makes it easily understandable for non-experts. However, depending on the specific process, other techniques could be applied as well; for a general discussion of clustering techniques, see [31].

3.3 Time Series Sensor Data Perspective

In TS analysis, [1] distinguishes three categories of techniques to cluster whole TS: distance-based features, using measures such as Euclidean or dynamic time warping (DTW) distance [30]; structure-based features, which characterise the whole TS; and shape-based features, created by searching for common motifs.

We use DTW distance, which allows a direct comparison of whole TS and is suitable for TS that are expected to share a common general structure as is the case in most manufacturing processes but can differ in length and speed (i.e. certain subsequences can last longer in one TS than in the other). Intuitively, it corresponds to the distance remaining between two series after eliminating timing differences, i.e., correcting for varying activity duration. It relies on the computation of a warping function mapping time points from two series together to minimise the distance between the two series. More specifically, given two series $A = a_1, a_2, ..., a_i, ..., a_n$ and $B = b_1, b_2, ..., b_j, ..., b_m$, with distance $d_{i,j} = ||a_i - b_j||$, DTW aims at finding an optimal mapping function $F = c_1, c_2, ..., c_k, ..., c_l$ such that the total distance $E(F) = \sum_{k=1}^{l} d(c(k)) \cdot w(k)$ is minimised:

$$DTW(A, B) = \min_{F} \left[\frac{\sum_{k=1}^{l} d(c(k)) \cdot w(k)}{\sum_{k=1}^{l} w(k)} \right] \tag{2}$$

where $w(k)$ is a weight coefficient for the elements of the mapping function.

Applying this for each pair of batches yields a distance matrix which can be used as input for clustering techniques like K-medoids or hierarchical clustering.

3.4 Multi-perspective Clustering

Once process instances are clustered separately in each perspective, the results are combined by clustering them together. Single-perspective clusters can be represented in different ways, such as using their labels as categorical features or computing distances to the centroids. We follow the latter approach, which retains more information for multi-perspective clustering.

Moreover, perspectives can be weighted to adjust their contribution to the multi-perspective clustering. For example, control-flow can be given more weight to ensure it has sufficient influence on the final clustering. Weights can also be used to account for differences in the number of clusters generated by each perspective, where more clusters may result in more features and a greater impact on the final clustering.

4 A Case Study in Smart Manufacturing

4.1 Use Case

Process Description. We applied TROPIC to a real use case at a partner company active in the production of chemical products. Their production process can be summarised in four main steps:

1. Preparing raw material and loading the tank;
2. Mixing the raw material in the tank;
3. Circulating the product through filters to remove impurities;
4. Bottling and packing the finished product.

Sometimes, the quality of the product is not high enough after filtering, i.e., some characteristics of the product do not meet the specifications. In this case, an adjustment is applied by loading additional raw materials into the tank and repeating steps two and three, resulting in the high-level production process depicted in Fig. 2.

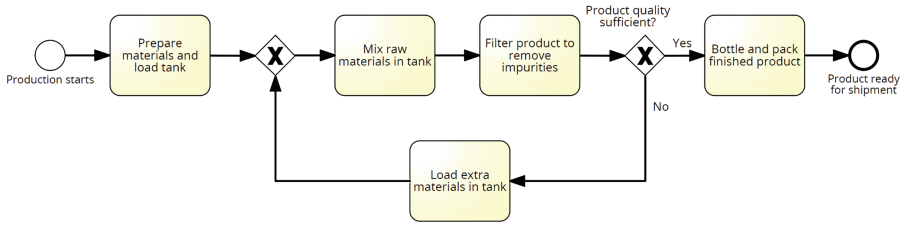


Fig. 2. High-level model of the process analysed in the experiment.

This seemingly simple process has to be executed with extreme precision and care as the slightest presence of impurities in the finished product greatly diminishes its quality. This is why the company is interested in analysing production logs and TS sensor data together to find out variation in process execution.

Data. Two main data sources are used in this use case: 1) logs from the production system, which contain the sequences of activities executed for each process instance and trace attributes and 2) TS data from sensors tracking the flow of the product in the four tanks and in the pipes leading through the filters every second. The data span a period from October 2020 to April 2022, representing 161 complete process instances and 199.4 million rows of sensor data.

Data Preprocessing. First, relevant TS pump circulation flow data was extracted for each batch. The data were resampled to one measurement per minute for smoothing and to reduce their length (the raw TS for the longest batch counted more than one million measurements before resampling), and some missing values due to the storage format were imputed. Finally, all data were normalised.

4.2 Clustering and Evaluation Approach

Multi-perspective Trace Clustering. We applied our two-step multi-perspective trace clustering approach to the obtained data. For the *control-flow perspective*, we followed a distance-based approach by computing the DL distance between the event sequences for each pair of batches and using the resulting distance matrix as input for the K-medoids. The number of clusters was set to five by plotting inertia and following the elbow method. The clusters contained 28, 23, 52, 22 and 36 instances, respectively. Secondly, regarding the *trace attributes perspective*, we applied the K-means algorithm with $K = 5$ (based on the elbow method). This yielded clusters of 23, 48, 41, 25 and 24 instances. Note that the attributes "tank open time" and "time in tank" are considered trace attributes as they measure batch quality and not timeliness. Third, we applied a distance-based TS clustering approach for the *TS sensor data perspective*, computing the DTW distance between the TS of each pair of batches to obtain a TS distance matrix used as input for K-medoids, with $K = 6$ (based on the elbow method), which formed clusters of sizes 9, 44, 59, 20, 21 and 8. Finally, to perform *multi-perspective clustering*, we computed the distances to centroids for each single-perspective clustering. Then we weighed the clusterings to take into account the different values of K in each clustering and applied K-means to all distances to centroids together, with $K = 4$ based on the elbow method. When K-means were applied, centroids initialisation was optimised to speed up convergence of the clustering by sampling centroids based on marginal inertia decrease, while when K-medoids were applied, medoids were randomly initialised.

Clustering Evaluation. The evaluation of clustering results is a challenging task that often depends on the specific domain and task at hand. A range of metrics are available to score clusterings based on intrinsic properties, such as the Davies-Bouldin (DB) score [6], which measures the similarity of clusters to their respective most similar cluster (lower value is better), or the Silhouette index [29], which compares the similarity between an instance and instances in its cluster with the similarity between this instance and instances in other clusters (higher value is better). Other metrics compare clusterings with known classes in the data or other clusterings, such as the Rand index [26], entropy, or purity. However, it is worth noting that better-formed clusters may not necessarily be more useful in practise, hence obtaining external validation from experts is critical to evaluate clustering results.

In our case study, we compared the clusters obtained from the multi-perspective approach with those derived from single-perspective clustering, using

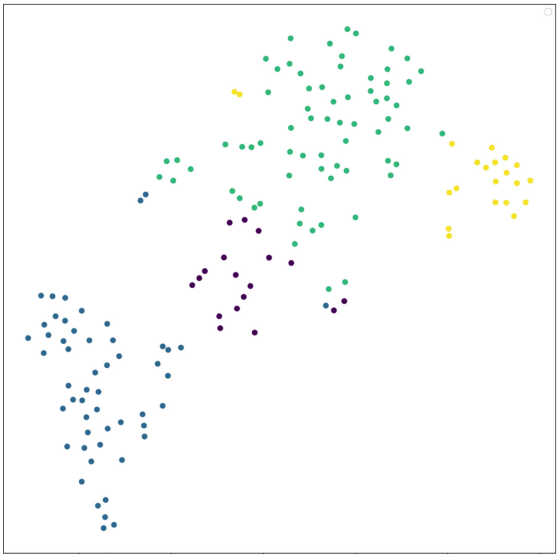


Fig. 3. Visualisation of multi-perspective clustering with t-SNE (cluster 1 = purple, cluster 2 = blue, cluster 3 = green, cluster 4 = yellow). (Color figure online)

both metrics and expert feedback. We computed silhouette indexes and DB scores for each clustering to assess the quality of the clusters in each approach. We also computed adjusted Rand indexes (ARI; where a higher value indicates higher similarity) and entropy scores (where a lower value indicates higher similarity) to determine the degree of similarity between the clusterings and to identify which perspective has the most influence on multi-perspective clustering. To validate our clustering results, we presented them to a senior process engineer at our partner company. Specifically, we showed the engineer the centroids of each multi-perspective cluster, as well as an overview of each cluster, including a directly-follows graph (DFG) for the control-flow, the mean or mode of trace attributes, and the DTW barrycenter average (DBA) [25] for the TS perspective, which is a method to compute the average of several TS taking into account potential time shifts.

4.3 Results

Multi-perspective clustering with K = 4 resulted in clusters of sizes 18, 53, 69, and 21 (see Fig. 3). In the remainder of this section, we provide visualisations of the clusters and report the values of the metrics and the interpretation and evaluation of the clusters by the process expert for each perspective.

Clustering Quality Assessment and Visualisation. The Silhouette score and the DB index are reported in Table 1. As can be seen, multi-perspective clustering has better scores than other clusterings for both metrics. Trace attributes

clustering has the worst scores, while control-flow and TS clusterings have similar values.

Table 1. Internal validation metrics for each clustering.

Metric	Multi-perspective	Control-flow	Trace attributes	Time series
Silhouette index	0.2516	0.0331	-0.0168	0.0284
DB score	1.3845	3.0526	4.5942	3.2061

Table 2 reports the cluster similarity metrics. Both entropy and ARI show that multi-perspective and control-flow clusterings have the highest similarity, i.e., they most often group the same instances together. On the other hand, trace attribute data clustering has high entropy and low ARI for all other clusterings, indicating that it forms very different clusters than the other perspectives.

We visualised the multi-perspective clusters by modelling the DFGs of their control-flows (see Figs. 4–5, where high-level steps from Fig. 2 are highlighted), computing the mean and the mode of their attributes (see Table 3) and plotting the DBAs of their TS data (see Figs. 6–7). DFGs and DBAs were used and are put forward in this paper as they can provide intuitive visualisations of the control-flow and the TS data of many instances of a process at once, enabling business experts to quickly understand and analyse whole clusters. Note that while all the results of the multi-perspective clustering are shown, only particularly interesting results are displayed for the other clusterings, and that activity labels as well as some trace attribute values were anonymised on request of the company.

Expert-Based Validation. When showing the multi-perspective clusters, the process expert categorised them as follows. Cluster 3, the largest cluster and the ones with the fewest distinctive characteristics, was identified as representing the standard execution of the process. Cluster 2 typically included traces with fewer adjustment activities and a lower material adjustments attribute than those in the other clusters, as shown in Fig. 4b and Table 3. In contrast, cluster 1 represented batches that required more adjustment activities and have a higher value for the material adjustments attribute (see Fig. 4a and Table 3) than batches in the other clusters. Having more adjustments also caused the filtering step to last longer, which can also be seen in the TS data by comparing Figs. 6a and 6b (filtering being characterised by long periods with a stable flow). Finally, cluster 4 included traces with missing activities that were necessary for proper process execution. These instances were identified as anomalies caused by improper logging of these activities.

4.4 Comparison of the Clusterings

In general, single-perspective clusters are more difficult to interpret than multi-perspective clusters. While control-flow clustering also groups together batches

Table 2. Pairwise similarity metrics values.

	Multi-perspective	Control-flow	Trace attributes	Time series
Multi-perspective entropy	0	0.8661	1.1585	0.9749
Multi-perspective ARI	1	0.2552	0.0301	0.0886
Control-flow entropy	1.1806	0	1.4154	1.3783
Control-flow ARI	0.2552	1	0.0340	0.0359
Trace attributes entropy	1.4790	1.4213	0	1.4663
Trace attributes ARI	0.0301	0.0340	1	0.0141
Time series entropy	1.2930	1.3817	1.4638	0
Time series ARI	0.0886	0.0359	0.0141	1

that required more adjustments, no cluster groups instances with fewer adjustments as neatly as multi-perspective cluster 2 (see Figs. 5a–5b). It is particularly difficult to recognise consistent patterns across perspectives in data clusters, while TS clusters succeed to some extent in grouping together instances with similar control-flows. Next to this, the most difficult perspective to interpret in all clusterings seems to be the TS perspective, where DBAs have difficulty capturing typical TS shapes, partly due to the presence of batches with missing data. This being said, DBAs based on TS clustering (see Fig. 7) seem more distinct and more easily interpretable.

5 Discussion

TROPIC successfully integrates TS sensor data in multi-perspective trace clustering, resulting in clusters that consider different process perspectives. The two-step structure makes it easy to disentangle the different perspectives, adjust their importance, and compare them. In our manufacturing use case, comparing multi-perspective trace clustering with single-perspective clustering showed that by leveraging underlying relationships between different perspectives, multi-perspective trace clustering could outperform single-perspective clustering even in their own perspective. For instance, multi-perspective trace clustering grouped instances with few adjustments better than control-flow clustering, as other perspectives helped recognise these instances.

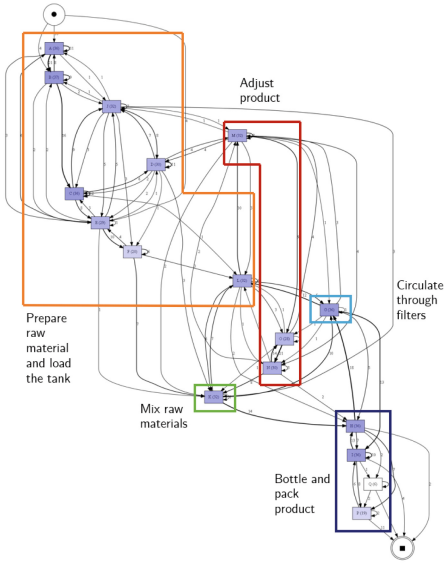
In addition, the process expert found multi-perspective clusters more meaningful from a business point of view, as they identified variants and anomalies. This insight could help the company investigate the differences between clusters 1 and 2 to reduce the number of necessary adjustments in the future.

Furthermore, some anomalous process instances were detected in the use case, although we did not apply any anomaly detection technique. This observation highlights the potential of multi-perspective anomaly detection using TROPIC by applying outlier detection to the distances to centroids.

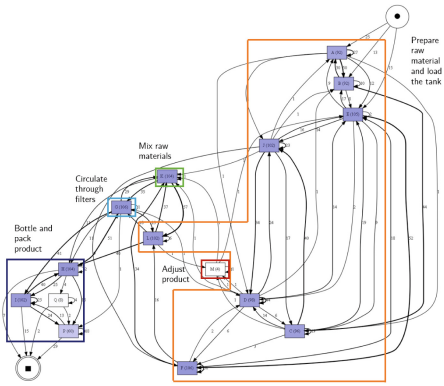
In addition, the choice of K for K-means and K-medoids clustering could have a great impact on the results of clustering at both stages. In this paper,

Table 3. Mean or mode of the trace attributes for each cluster of all clusterings (standard deviations between brackets).

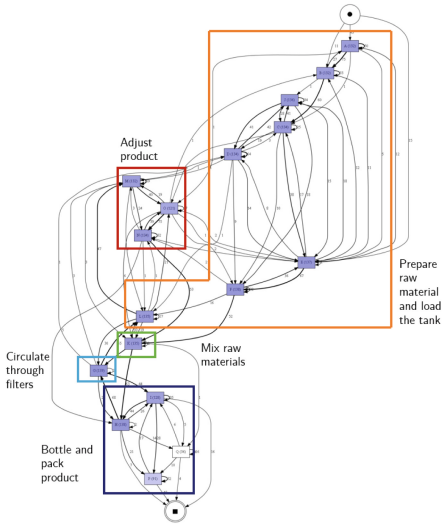
Cluster	Materials	Adj. materials	Tank open time	Solvent	Time in tank	Tank
Multi perspective 1	5.7222 (5.3776)	3.8889 (3.2519)	6392.9444 (4161.9357)	A /	381671.8889 (223893.4571)	T3 /
Multi perspective 2	7.2453 (2.0466)	0.0755 (0.3848)	7552.4151 (2760.7288)	B /	261959.6038 (60081.1298)	T4 /
Multi perspective 3	7.0 (1.4349)	2.7681 (1.5449)	8689.6377 (2785.0268)	B /	298360.3478 (84948.1399)	T1 /
Multi perspective 4	7.2381 (1.9211)	2.8095 (1.4703)	10599.1429 (2919.2145)	B /	260751.6667 (97850.3688)	T4 /
Control-flow 1	7.8929 (3.6952)	1.1071 (2.3308)	7877.0 (1878.0679)	B /	270904.6786 (69923.7843)	T4 /
Control-flow 2	6.2609 (2.4349)	3.9565 (2.5132)	8857.0 (3929.4461)	A /	325021.1739 (118121.8841)	T1 /
Control-flow 3	7.1538 (1.9742)	2.0769 (1.4799)	8587.3077 (3264.1794)	B /	301473.25 (108380.1455)	T4 /
Control-flow 4	5.8636 (2.1447)	2.2273 (1.3778)	9019.5909 (3878.9293)	A /	283188.4545 (118979.1435)	T4 /
Control-flow 5	7.1111 (1.6695)	1.25 (1.9911)	7452.2222 (2711.0095)	A /	273584.1389 (124958.7138)	T1 /
Trace attributes 1	8.4348 (3.4089)	4.6957 (2.6187)	9494.7391 (3112.0703)	B /	291919.2609 (94352.3756)	T4 /
Trace attributes 2	7.2083 (1.688)	1.1667 (1.3262)	8225.375 (2075.6587)	B /	257210.7083 (65348.9021)	T4 /
Trace attributes 3	6.7317 (2.3667)	1.3415 (1.5266)	7048.2683 (3339.7471)	A /	282512.2683 (73998.1907)	T1 /
Trace attributes 4	5.84 (2.5113)	2.52 (1.8735)	7396.36 (3202.4594)	A /	354616.76 (180006.1447)	T3 /
Trace attributes 5	6.6667 (2.1196)	1.75 (1.7508)	10434.7083 (3427.8528)	B /	304496.7917 (128730.3709)	T4 /
TS 1	5.6667 (2.2913)	2.2222 (1.8559)	9548.1111 (4358.0709)	Other /	217580.6667 (38620.8273)	T4 /
TS 2	6.6136 (2.4133)	1.75 (1.6999)	7262.3182 (2992.4131)	A /	355988.5455 (123518.9904)	T3 /
TS 3	7.1356 (2.5492)	2.0169 (2.4878)	8272.678 (2755.9548)	B /	260759.0508 (97720.4561)	T4 /
TS 4	7.7 (1.8382)	1.9 (1.8035)	9000.9 (2929.1435)	B /	267001.65 (68154.5069)	T4 /
TS 5	6.7143 (2.1941)	2.5714 (2.0874)	9366.0 (4099.0399)	B /	312286.619 (120953.309)	T2 /
TS 6	8.0 (3.4641)	2.0 (2.3299)	8406.375 (2185.4744)	B /	239005.5 (24086.8173)	T4 /



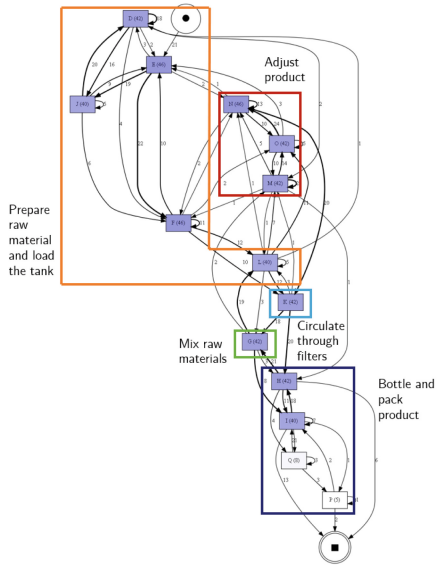
(a) DFG for cluster 1.



(b) DFG for cluster 2.



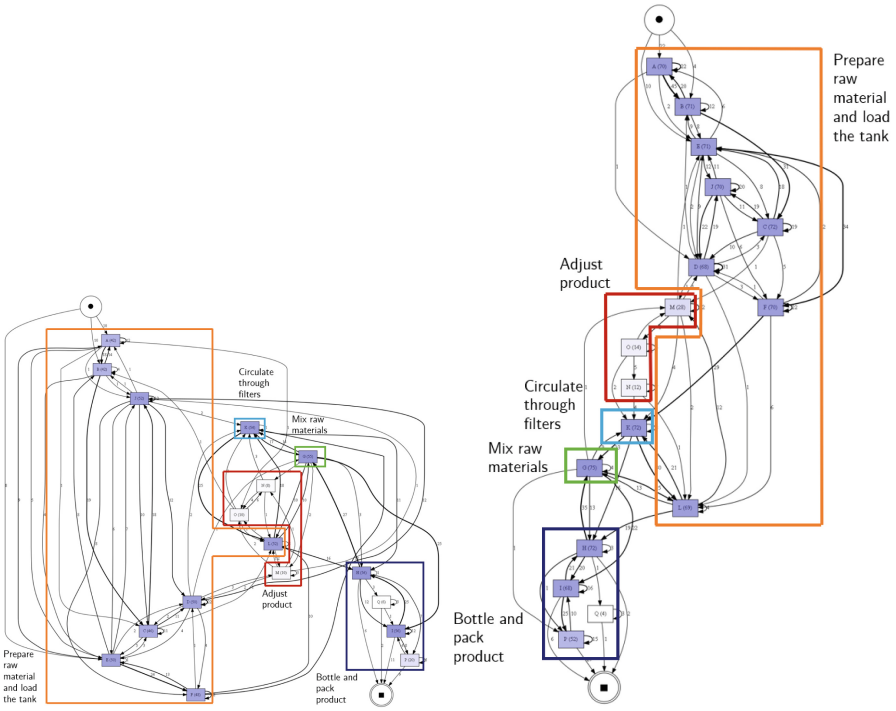
(c) DFG for cluster 3.



(d) DFG for cluster 4.

Fig. 4. DFGs for each cluster of the multi-perspective clustering.

the popular elbow method was used and yielded good results, as the clusters formed were insightful from a business perspective. Future work could investigate more complex methods to determine the value of K, e.g., based on stability or separation as in [7].

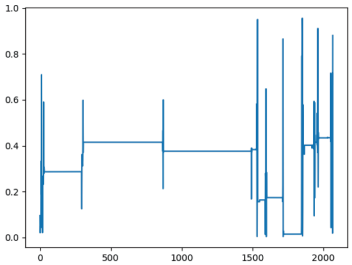


(a) DFG for cluster 1. (b) DFG for cluster 5.

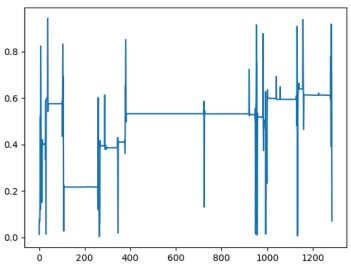
Fig. 5. DFGs for clusters 1 and 5 of the control-flow clustering.

However, ARI and entropy indicated that the control-flow perspective produced a clustering more similar to the other perspectives. This result suggests that the control-flow perspective might be more important than other perspectives in the multi-perspective trace clustering. Weighting the perspectives could rebalance their contributions, but as all perspectives are correlated, weighting may not fundamentally change the clustering in the use case.

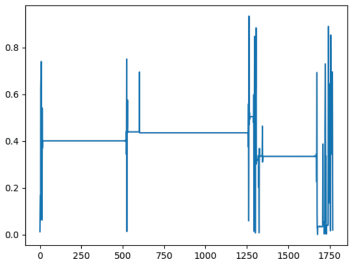
Finally, although we focused on three specific perspectives in this paper, we believe our approach could be extended to consider other perspectives. For example, a similar approach to that applied to the TS data obtained from IoT sensors could be applied to other processes that evolve over time, such as process performance. Such a different perspective could serve as a substitute for one of the current three dimensions, or the approach could easily be adapted to a higher dimensionality, allowing for several other perspectives to be included, such as the resource perspective.



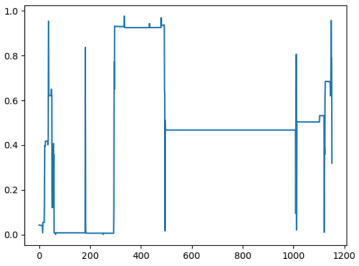
(a) DBA for cluster 1.



(b) DBA for cluster 2.

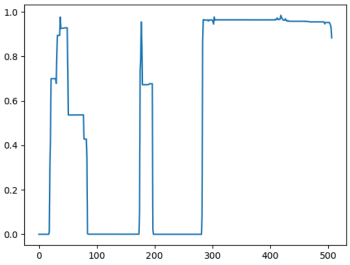


(c) DBA for cluster 3.

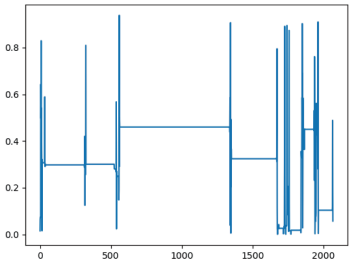


(d) DBA for cluster 4.

Fig. 6. DBAs for each cluster of the multi-perspective clustering.



(a) DBA for cluster 1.



(b) DBA for cluster 5.

Fig. 7. DBAs for clusters 1 and 5 of the TS clustering.

6 Conclusion

In this paper, we presented a novel approach for multi-perspective trace clustering of manufacturing processes that considers three perspectives: control-flow, trace attributes, and TS sensor data. This approach can reveal process variants that are homogeneous across all three perspectives simultaneously. We evaluated the approach in a real-life use case of a smart manufacturing process, where it

revealed meaningful clusters and anomalous instances for a specific IoT-enhanced BP, both actionable insights to improve process design and execution.

In future work, we plan to extend this approach in various ways. One possibility is to propose a generalisation to n arbitrary perspectives. We could also consider including event attributes and incorporating TS data at the event level. Furthermore, we could explore other clustering techniques for the multi-perspective clustering if our approach were to be used for more flexible types of processes, such as ensemble clustering methods or soft clustering techniques. Finally, we find that integrating contextual information in the log in the form of events, as suggested in [3], is an interesting alternative approach.

Acknowledgement. This research was supported by the Flemish Fund for Scientific Research (FWO) with grant number G0B6922N.

References

1. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y.: Time-series clustering-a decade review. Inf. Syst. **53**, 16–38 (2015)
2. Banham, A., Leemans, S.J., Wynn, M.T., Andrews, R., Laupland, K.B., Shinners, L.: xPM: enhancing exogenous data visibility. Artif. Intell. Med. **133**, 102409 (2022)
3. Bertrand, Y., De Weerdt, J., Serral, E.: A bridging model for process mining and IoT. In: Munoz-Gama, J., Lu, X. (eds.) ICPM 2021. LNBIP, vol. 433, pp. 98–110. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98581-3_8
4. Bertrand, Y., De Weerdt, J., Serral, E.: Assessing the suitability of traditional event log standards for IoT-enhanced event logs. In: In: Cabanillas, C., Garmann-Johnsen, N.F., Koschmider, A. (eds.) Business Process Management Workshops. BPM 2022. Lecture Notes in Business Information Processing, vol. 460, pp. 63–75. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25383-6_6
5. Bose, R.J.C., Van der Aalst, W.M.: Context aware trace clustering: towards improving process mining results. In: Proceedings of the 2009 SIAM International Conference on Data Mining, pp. 401–412. SIAM (2009)
6. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **2**, 224–227 (1979)
7. De Koninck, P., De Weerdt, J.: Similarity-based approaches for determining the number of trace clusters in process discovery. In: Koutny, M., Kleijn, J., Penczek, W. (eds.) Transactions on Petri Nets and Other Models of Concurrency XII. LNCS, vol. 10470, pp. 19–42. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-662-55862-1_2
8. De Weerdt, J.: Trace clustering (2019)
9. De Weerdt, J., Vanden Broucke, S., Vanthienen, J., Baesens, B.: Active trace clustering for improved process discovery. IEEE TKDE **25**(12), 2708–2720 (2013)
10. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd, vol. 96, pp. 226–231 (1996)
11. Fani Sani, M., van Zelst, S.J., van der Aalst, W.M.P.: Applying sequence mining for outlier detection in process mining. In: Panetto, H., Debruyne, C., Proper, H.A., Ardagna, C.A., Roman, D., Meersman, R. (eds.) OTM 2018. LNCS, vol. 11230, pp. 98–116. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02671-4_6

12. Fulcher, B.D., Little, M.A., Jones, N.S.: Highly comparative time-series analysis: the empirical structure of time series and their methods. J. R. Soc. Interface **10**(83), 20130048 (2013)
13. Greco, G., Guzzo, A., Pontieri, L., Saccà, D.: Mining expressive process models by clustering workflow traces. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 52–62. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24775-3_8
14. Grüger, J., Geyer, T., Kuhn, M., Braun, S.A., Bergmann, R.: Verifying guideline compliance in clinical treatment using multi-perspective conformance checking: a case study. In: ICPM Workshops, pp. 301–313 (2021)
15. Jablonski, S., Röglinger, M., Schönig, S., Wyrtki, K.M.: Multi-perspective clustering of process execution traces. EMISAJ **14**, 2 (2019)
16. Janiesch, C., et al.: The internet of things meets business process management: a manifesto. IEEE Systems, Man, Cybern. Mag. **6**(4), 34–44 (2020)
17. Kohonen, T.: The self-organizing map. Proc. IEEE **78**(9), 1464–1480 (1990)
18. Leno, V., Dumas, M., Maggi, F.M., La Rosa, M.: Multi-perspective process model discovery for robotic process automation. In: CAiSE 2018, vol. 2114, pp. 37–45 (2018)
19. Luengo, D., Sepúlveda, M.: Applying clustering in process mining to find different versions of a business process that changes over time. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM 2011. LNBIP, vol. 99, pp. 153–158. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28108-2_15
20. Lull, J.J., et al.: Exploration with process mining on how temperature change affects hospital emergency departments. In: Leemans, S., Leopold, H. (eds.) ICPM 2020. LNBIP, vol. 406, pp. 368–379. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72693-5_28
21. MacQueen, J.: Classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California, Los Angeles (1967)
22. Mannhardt, F.: Multi-perspective process mining. In: BPM (Dissertation/Demos/Industry), pp. 41–45 (2018)
23. Mannhardt, F., De Leoni, M., Reijers, H.A., Van Der Aalst, W.M.: Balanced multi-perspective checking of process conformance. Computing **98**, 407–437 (2016)
24. Mannhardt, F., de Leoni, M., Reijers, H.A., van der Aalst, W.M.P.: Data-driven process discovery - revealing conditional infrequent behavior from event logs. In: Dubois, E., Pohl, K. (eds.) CAiSE 2017. LNCS, vol. 10253, pp. 545–560. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59536-8_34
25. Petitjean, F., Ketterlin, A., Gançarski, P.: A global averaging method for dynamic time warping, with applications to clustering. Pattern Recogn. **44**(3), 678–693 (2011)
26. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
27. Rdusseeun, L., Kaufman, P.: Clustering by means of medoids. In: Proceedings of the Statistical Data Analysis Based on the L1 Norm Conference, vol. 31 (1987)
28. Rodriguez-Fernandez, V., Trzcionkowska, A., Gonzalez-Pardo, A., Brzychczy, E., Nalepa, G.J., Camacho, D.: Conformance checking for time-series-aware processes. IEEE TII **17**(2), 871–881 (2020)
29. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
30. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE TASSP **26**(1), 43–49 (1978)

31. Saxena, A., et al.: A review of clustering techniques and developments. Neurocomputing **267**, 664–681 (2017)
32. Scheibel, B., Rinderle-Ma, S.: Online decision mining and monitoring in process-aware information systems. In: Ralyte, J., Chakravarthy, S., Mohania, M., Jeusfeld, M.A., Karlapalem, K. (eds.) Conceptual Modeling. ER 2022. Lecture Notes in Computer Science, vol. 13607, pp. 271–280. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17995-2_19
33. Seiger, R., Franceschetti, M., Weber, B.: An interactive method for detection of process activity executions from IoT data. Future Internet **15**(2), 77 (2023)
34. Seiger, R., Zerbato, F., Burattin, A., García-Bañuelos, L., Weber, B.: Towards IoT-driven process event log generation for conformance checking in smart factories. In: EDOCW 2020, pp. 20–26. IEEE (2020)
35. Trzcionkowska, A., Brzychczy, E.: Practical aspects of event logs creation for industrial process modelling. MAPE **1**(1), 77–83 (2018)
36. Valderas, P., Torres, V., Serral, E.: Modelling and executing IoT-enhanced business processes through BPMN and microservices. J. Syst. Softw. **184**, 111139 (2022)
37. Valencia Parra, Á., Ramos Gutiérrez, B., Varela Vaca, Á.J., Gómez López, M.T., García Bernal, A.: Enabling process mining in aircraft manufactures: extracting event logs and discovering processes from complex data. In: BPM2019IF (2019)
38. Ward, J.H., Jr.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)