Resource-Centric Process Mining: Clustering Using Local Process Models

Landelin Delcoucq , Fabian Lecron , Philippe Fortemps
Department of Engineering Innovation Management, Faculty of Engineering, University of Mons, Mons, Belgium

Wil.M.P. van der Aalst
Process and Data Science (PADS), RWTH Aachen University, Aachen, Germany

ABSTRACT

In this paper, we focus on the resource perspective in the context of process mining. Most process mining techniques focus on the control-flow to uncover problems related to performance or compliance. However, the behavior of resources (e.g., employees) influences the effectiveness and efficiency of processes and should not be considered as secondary. We aim to identify resources exhibiting similar behavioral patterns that go beyond just looking at the mix of activities performed. We want to be able to identify subgroups of resources that perform similar activities but in a different order. We also provide a comparison between existing ways of grouping resources into roles and our resource-centered approach that takes into account the order in which work is performed. We will compare the results of clustering based only on the activities performed and clustering based on local process models that identify work patterns. Experiments are considered on synthetic and real data.

CCS CONCEPTS

• **Information systems** → **Clustering**; • **Computing methodologies** → *Artificial intelligence*; • **Social and professional topics** → Medical records.

KEYWORDS

process mining, clustering, local process model

ACM Reference Format:
Landelin Delcoucq , Fabian Lecron , Philippe Fortemps and Wil.M.P. van der Aalst. 2020. Resource-Centric Process Mining: Clustering Using Local Process Models. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3341105.3373864

1 INTRODUCTION

The process mining aims to distill a structured description of a process based on a set of real executions of it [10]. Those executions are described by structured information called *logs*. The process mining includes the following three main types of analysis [8]: discovery, conformance and enhancement.

The main focus of process mining is often on the control-flow perspective. However, over time several authors have worked on adding a more human-centered resource perspective [6, 9]. The organizational perspective, also referred to as resource perspective, aims to describe what happens behind the process, how are structured the organizations and what are the interactions between their members. One of the main questions related to this perspective is the identification of resources exhibiting similar behaviors. This problem is the main purpose of this paper: identifying clusters of resources sharing common behaviors, those behaviors being described by local process models [7].

At the current time, there are two main ways used to deal with this issue. One approach, based on social networks [9], measures how two resources are linked by the similarity of their activities (e.g., similar tasks and handover-of-work). An intuitive proposed solution is to put together people sharing the same kinds of links [1, 9]. An alternative approach is to cluster traces [2, 5, 9]. Both of these approaches are able to group resources sharing activities or simple sequences of activities. However, they are not able to identify common sub-processes (i.e. activity patterns) or they need organizational background knowledge [4] .

The approach presented in this paper proposes an adaptation of the trace clustering based on the local process models [7]. Such local process models represent frequent sub-processes and, in the context of this paper, they refer to recurring behaviors performed by several resources. Those behaviors will be called *tasks* (i.e., tasks refer to groups of activities performed for one or more cases). The current approaches only consider the resources as activity performers and the key point of our approach is to consider them as task performers. This higher level of granularity will allow us to provide an improved representation of the process.

Concretely, let us assume a healthcare group composed of several hospitals. Two of them cure the patients in the same way, they perform the same processes and they have one more common point: their nurses work similarly. In both hospitals, there are two kinds of employees. Some of them work sequentially, they take care of one patient from its admission to the end of the hospitalization then take charge of another patient. The other nurses work in batch, they prefer to perform the same activity on a set of patients before moving to another activity. To improve its organization, the healthcare group would like to identify (and then regroup) the nurses working in the same way in the same hospital. Current approaches are unable to distinguish employees performing the same set of activities. Our approach based on the *tasks* (i.e., groups of activities) will lead to a better grouping of resources. The *tasks*, patterns of

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/10.1145/3341105.3373864

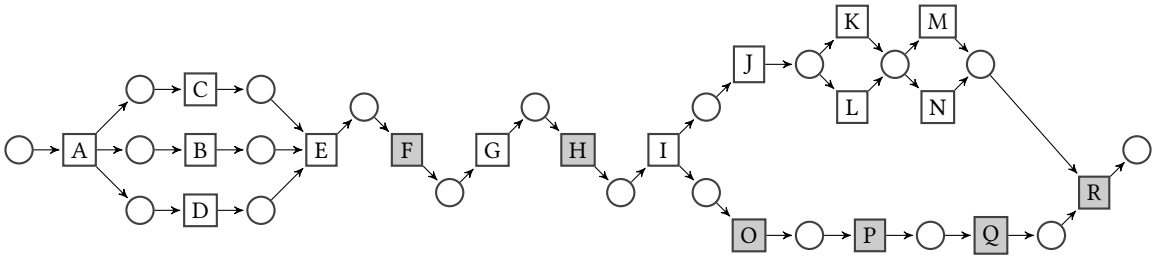


Figure 1: Running example process (A: admission, B: blood test, C: blood pressure test, D: allergological tests, E: collecting results, F: analyzing results, G: preoperative test, H: anesthesia, I: medical equipment checking, J: anesthesia monitoring, K: injection type 1, L: injection type 2, M: wound cleaning, N: blood suction, O: incision, P: surgical procedure, Q: stitching, R: postoperative checking). The white activities are performed by nurses and the grey activities are performed by doctors.

successive activities, are completely different for someone working sequentially and someone working in batch.

This paper aims to show that grouping resources based on the activities they perform, fails to capture significant differences in behavior. Therefore, we group resources based on differences in behavior characterized by local process models. For this purpose both approaches are compared on synthetic and a real-life dataset. The remainder of the paper is organized as follows. The next section defines the key points of the proposed approach (point of view of resources and local process mining) and introduces the basic concepts of clustering methods. Sections 3 and 4 describe the experimental approach developed to support the proposed hypothesis. Section 5 talks about the limitation of the approach and its future perspectives. Finally, the last section concludes the paper.

2 LPM RESOURCE CLUSTERING APPROACH

The proposed approach and the concepts used will be explained using a running example. This example is introduced in Subsection 2.1. In Subsection 2.2, we introduce the need for a different kind of traces, i.e. resource traces. Subsection 2.3 explains the extraction of tasks (by means of local process models). Subsection 2.4 describes the clustering stage.

2.1 Running Example

We consider a surgical process, consisting of the 18 medical acts each patient undergoes (Figure 1). The activities in white are performed by nurses. The gray activities are performed by doctors.

2.2 Creating Traces from Event Data

Process mining starts from event logs where events have at least four attributes: case, activity, resource, and timestamp. Table 1 shows an example event log. The activity attribute refers to the action performed. The term *case (ID)* is used to define the entity over which the process is performed. The last attribute is the timestamp representing when an *activity* is performed.

Table 1: Partial logs of the running example

Case ID	Activity	Timestamp	Resource
Patient 1	Admission (A)	01/01/18 09:04:27	Nurse John
Patient 5	anesthesia (H)	01/01/18 11:41:04	Nurse Tim
Patient 5	Injection type 1 (K)	01/07/18 08:12:33	Nurse Tim
Patient 2	Surgical procedure (P)	01/08/18 15:14:59	Doctor Jane
Patient 3	Analyzing results (F)	01/12/18 10:32:01	Doctor Jane
Patient 4	Incision (O)	01/21/18 14:00:52	Doctor Tom
Patient 5	Wound cleaning (M)	02/03/18 16:58:00	Nurse Tim
...	...	...	...

Usually, a trace describes the sequence of activities performed for a given case. In our example, it describes all the medical acts performed on a patient. The model extracted from those traces is depicted in Figure 1 which represents the process performed by a patient.

Since our approach is resource-centric, we need to consider event data from a new point of view. Therefore, we introduce a new way to build the traces from the logs. Resource traces represent the sequence of activities performed by a resource. In other words, the resource is chosen as the identifier and the activity as the trace attribute, and we sort the items by their timestamps. In our example, it represents all the activities performed by a particular nurse or a particular doctor.

For the given example, the trace corresponding to the resource *Nurse Tim* is

$$\text{anesthesia(H)}_{patient5} \text{ - Injection type 1(K)}_{patient5} \text{ - Wound Cleaning(M)}_{patient5}$$

Based on those traces, a preliminary conclusion can already be drawn, the resource traces have a structure different from the traces built using the case/activity way. A classical process, even if it may have loops, is built following a global linear shape because the processes have a precise start-point and a precise end-point. Patient always starts the process with *Admission* and ends it with the *Postoperative Checking*. This assertion is no longer valid concerning the traces built using the resources.

To fully understand and identify the behavior of the resources, the notion of a case (the patient) must be added. The **activities**

printed in bold face, in the traces, mean that the resource starts to work on a different patient. This representation provides new insights and allows us to identify different behaviors e.g. Nurse John performed A - B - C - D - E - G - I - J for the same patient. Nurse Sarah performed A - B for one patient and A - B - C - D - G - I for another patient (Table 2). The synthetic dataset, exposed later in this paper is based on the same process and will be performed by resources behaving in the same way.

Table 2: Traces built on the logs and using the resource/activity view (note that an activity name is printed in boldface when it is for a new case)

Resources	Traces
Nurse John	**A** - B - C - D - E - G - I - J
Nurse Sarah	**A** - B - **A** - B - C - D - G - I
Nurse Tony	**A** - **A** - B - D - C - **A** - B - C
Nurse Willy	**A** - B - C - D - **B** - D - **C** - **B**
Nurse Eve	**A** - B - **B** - D - **A** - B - D - C
Nurse Tim	**G** - **G** - **I** - **I** - **J** - **J** - K - M
Nurse Suzan	**I** - **I** - **J** - **J** - K - **L** - M - **M**
Nurse Sue	**A** - B - **G** - **E** - **I** - **D** - **J** - **C**
Nurse Mary	**E** - **E** - **E** - **B** - **B** - **E** - **E** - **E**
Doctor Jane	**F** - H - O - P - **F** - H - O - P
Doctor Tom	**F** - **F** - H - **H** - O - **O** - P - **P**
Doctor Sam	**F** - **F** - **H** - **H** - **O** - **O** - **P** - **P**

2.3 Local Process Models

To extract the tasks from traces, as described in Subsection 2.2, we use the local process models. A local process model [7] is a local sub-process describing the most frequent behaviors of an event log. Those local models are able to represent loops, concurrency, choices and sequential patterns (i.e., all the elements that can be extracted from the data by the process mining algorithms).

From the five first resources of Table 2, the local process model algorithm allows us to extract the pattern depicted in Figure 2. This LPM models the behavior of the five first nurses: first A, then B and finally (C then D) or (D then C).

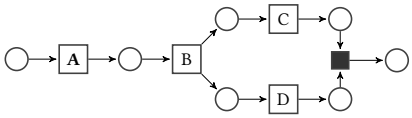


Figure 2: Local process model extracted from the five first traces of Table 2

2.4 Clustering

In the present work, we used the well-known clustering algorithm: KMeans. A first approach consists in using the KMeans algorithm on an activity-based representation of the resources. Each resource is described by the number of times each activity is performed by this resource (see the example in Table 3a). A second approach consists in using the same KMeans algorithm, but on the number of times each task (extracted by means of the LPMs) is performed by the resources. Table 3b shows the frequencies of two LPMs, the first is coming from Figure 2 and the second represents the subtrace E-E-E coming from the Nurse Mary.

Table 3: Representation by frequency vector

(a) Activity-based frequency vector

Resources	A	B	C	D	E	G	I	J
Nurse John	1	1	1	1	1	1	1	1
Nurse Sue	1	1	1	1	1	1	1	1
Nurse Mary	0	2	0	0	6	0	0	0

(b) LPMs frequency vector

Resources	A - B - C//D	E - E - E
Nurse John	1	0
Nurse Sue	0	0
Nurse Mary	0	2

From this example, the activity-based clustering is able to differentiate completely different resources (i.e. performing different activity subsets): Nurse Mary against Nurse John and Nurse Sue. On top of that, our approach, based on the LPMs, will be able to differentiate different kinds of nurses (in terms of behaviors). Indeed, the local process models represent recurring performed behaviors. Even if both Nurse John and Nurse Sue perform the same activities, they did it in a different way, they exhibit different behaviors represented by distinct LPMs.

3 COMPARISON OF THE APPROACHES BASED ON SYNTHETIC LOGS

To check the hypothesis that the LPMs can improve the clustering, we compare the results of the clustering algorithm using as input the performed activities (their frequencies) and using as input the performed LPMs (their frequencies). Both frequencies are extracted from the synthetic logs. We will analyze the results of the clustering regarding to the ground-truth to evaluate their correctness but also by comparing them directly each other.

3.1 Presentation of the Dataset

We use CPN Tools to create a synthetic dataset, simulating the surgical treatment of patients as described in Figure 1. The dataset focuses on activities (E, F, G, H, I, J, O, P, Q), in order to cover all the doctors and the group of nurses performing the central part of the process. Indeed, the first four activities corresponding to the admission are performed by completely different nurses and it's the same for the five last activities. When a patient enters the core process, a class of nurses is chosen for the care steps and the medical steps are performed by the doctor class (see also Table 4).

- Class N1: For each new patient, a free nurse cares for him from the start to the end. When the process is complete, the nurse cares for a new patient. Such nurses work sequentially.

Table 4: Classes examples

Class	Trace
N1	**E** - G - I - J - **E** - G - I - J
N2	**G** - I - **G** - **I** - **G** - J - **I** - **J**
N3	**E** - **E** - G - **G** - **I** - **I** - J - **J**
N4	**E** - **G** - I - **J** - **E** - G - **I** - **J**
D1	**F**- H - P - **H** - O - Q - **Q**- **P**

- Class N2: For each required activity, regardless of the patient, a free nurse performs it then becomes free again. The nurses work in an "as they come" way.
- Class N3: The nurses work in batch, a free nurse cares for two patients from the start to the end by performing the activities in a parallel way. Such a nurse performs the process sequentially, from the first activity to the last one, but with two identified patients at a time.
- Class N4: The nurses perform the activities in a sequential way but regardless of the patient.
- Class D1: For each required activity, regardless of the patient, a free doctor performs it then becomes free again.

The generated dataset considers the treatment of 300 patients. Each class (of nurses or of doctors) is a pool of 5 people. The dataset is built under the assumption that the workload is evenly distributed as depicted in Table 4. From the patient log, we extract the resource traces for each nurse and each doctor.

Three comparisons between the activity-based and the LPM-based approaches will be executed. The first will be based on the whole dataset and the purpose of the comparison is to identify the nurses and the doctors. The other comparisons will focus on the nurses. The second will use the LPMs regardless of the patients and the third will consider the patients in the generation of the LPMs.

3.2 Indicators of Comparison

To evaluate the results taking the ground-truth into account, there are two main indicators: the confusion matrix and the adjusted rand index. The confusion matrix is a two-dimensional matrix where the columns represent the classes and the rows the clusters. In this matrix, each instance is placed in the cell corresponding to its cluster and to its class, consequently the representation of the ground-truth is a diagonal matrix. Based on this matrix, the *recall* and the *precision* can be computed:

The *precision* related to class i is defined as: $Pre(i) = \frac{n_{ii}}{n_{.i}}$ where n_{ij} is the number of resources coming from the class i in cluster j. $n_{i.}$ represents all the resources coming from the class i and $n_{.i}$, all the resources in cluster i.

The *recall* related to class i is defined as: $Rec(i) = \frac{n_{ii}}{n_{i.}}$.

The global value for a particular clustering is obtained by averaging all the classes values.

The *adjusted Rand* index [3] is based on the ratio between the number of pairs of resources that are either in the same group or in different groups in the clustering and in the ground-truth divided by the total number of pairs of resources. The optimal value is then 1 and the worse value is 0.

3.3 Comparison between Nurses and Doctors

Performing the KMeans with only two clusters (k=2) allows to distinguish perfectly the nurses and the doctors. We obtain the same result independently of the approach used, i.e., activity-based and LPM-based clustering both separate nurses and doctors perfectly (Table 5).

Table 5: Confusion matrix: activity-based and LPM-based clustering

	Cluster 1	Cluster 2
Nurses	40	0
Doctors	0	10

3.4 Comparison between the Nurses Regardless of the Patient

To highlight the specific behaviors of the nurses, we now focus on Cluster 1 and try to subcluster it by means of KMeans with k=4. In this context, the LPM approach provides better results (see Table 6a and Table 6b). The clustering based on the activities cannot identify the difference between the nurses N1, N3 and N4 i.e. people performing the same activities in different ways cannot be identified.

Table 6: Confusion matrix for subclustering the nurses

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
N1	5	0	0	0
N2	1	2	1	1
N3	5	0	0	0
N4	5	0	0	0

(a) Activity-based clustering

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
N1	5	0	0	0
N2	0	3	1	1
N3	0	0	5	0
N4	5	0	0	0

(b) LPM clustering

By contrast, if we extract tasks by means of LPMs regardless of the patient id, the clustering already provides improved results. The proposed approach is able to distinguish nurses performing the same activities in a different way (sequential vs batch) however the approach is not able to classify correctly the instances coming from the class N1 and N4.

Based on the confusion matrices, it is clear that the recall and the precision of the proposed approach are better. The adjusted rand index confirms that the algorithm using the LPM classifies more instances similarly as the ground-truth.

Table 7: Recall, precision and adjusted rand index for both approaches

Indicator	activity-based clustering	LPM-based clustering
Precision	0.35	0.65
Recall	0.33	0.58
Adjusted Rand Index	0.0123	0.5162

3.5 Comparison between the Nurses Taking into Account the Patients

To completely model the behavior of the nurses, we take into account the patients and more precisely when the nurses start to perform activities on a new patient. The goal is to be able to differentiate classes N1 and N4. An activity (e.g., B) carried out by a given resource may apply either to the same patient as the previous activity of this resource or to another patient. In the first case, the activity will be represented by its letter in regular font (B); in the second case, it will be represented by the letter in boldface (B). Thus, in addition to the activities, we also have the indication of a possible change of patient between two activities. By means of LPMs, we extract new tasks with activities that may include a change patient. With such patient-aware LPMs, the four classes are correctly identified (Table 8).

Table 8: Confusion matrix: patient-aware LPM-based clustering

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
N1	5	0	0	0
N2	0	5	0	0
N3	0	0	5	0
N4	0	0	0	5

The previous approach based on the LPMs was patient-blind: it did not take the patient id into account. Thus, it was not able to cluster efficiently the class N2 (the nurses with the "as-they-come" behavior) (Table 6b). In this subsection, we add a kind of simple patient awareness, more precisely the awareness of a patient-change. This allows to enrich the set of possible LPMs. There can be different tasks according to the patient switching policies. With such an approach, the class N2 is now correctly clustered (Table 8). This issue will be more deeply discussed in Subsection 5.1

4 COMPARISON OF THE APPROACHES BASED ON REAL-LIFE LOGS

The assumption that the LPMs improve the clustering of the resources is supported for the synthetic dataset but we have to check it on a real-life dataset. Again, we compare the results of the clustering algorithm using as input the performed activities (their frequencies) and using as input the performed LPMs (their frequencies) but, this time, both frequencies are extracted from a real-life dataset. The indicators used for the synthetic dataset will be used to evaluate the result.

4.1 Presentation of the Dataset

The dataset is based on treatments coming from the radiation oncology department of a hospital over two years. The dataset consists of 3058 treatments containing 29019 activities performed by 30 resources. There are three kinds of resources performing different activities: 15 nurses, 9 physicists and 6 doctors. The initial value of the parameter k is then *3*. The distribution of the dataset is between the resources.

The radiation oncology has clear procedures and well-defined processes composed of 11 activities. These activities describe the preparation of the radiation therapy itself. There is a stage of simulation followed by a stage of computation of the parameters by the doctors and the physicists. There are also several activities concerning the coordination between the resources and the verification of the parameters.

4.2 Comparison between the Resources

Table 9 shows that approach based on the activities is not able to differentiate the different kinds of workers. Table 9a indicates that the majority of the instances are classified in Cluster 1 whereas Table 9b highlights that the LPM-based clustering is able to identify partly the physicist.

Table 9: Confusion matrix for the real-life dataset

	Cluster 1	Cluster 2	Cluster 3
Nurse	14	1	0
Doctor	6	0	0
Physicist	6	0	3

(a) Activity-based clustering

	Cluster 1	Cluster 2	Cluster 3
Nurse	15	0	0
Doctor	6	0	0
Physicist	3	5	1

(b) LPM clustering

Table 10 shows that the approach using the LPMs provides better results on all the indicators. However, comparatively to the synthetic dataset (Table 7), the improvement of the proposed approach on the quality of the clustering is weak. It can be explained by the way in which the dataset was built. Theoretically, the encoding of an activity in the log indicates the resource who has performed the corresponding activity. In practice, the activities performed by the doctors are regularly encoded by the nurses. It causes an overlap between the doctors and the nurses.

Table 10: Recall, precision and adjusted rand index for both approach on the full real-life dataset

Indicator	activity-based clustering	LPM-based clustering
Precision	0.42	0.51
Recall	0.51	0.54
Adjusted Rand Index	0.051	0.257

The purpose of the proposed approach is to take into account the links between the performed activities rather than the activities themselves. We know that the doctors and the nurses encode the same activities but not in the same way. Their behaviors are different and Table 11 proves the efficiency of our approach. On this table, the approach was used on the same dataset, without the physicist and with $k = 2$. The activity-based clustering is not able to differentiate the doctors and the nurses whereas the LPM-based clustering identifies the different behaviors and provides the right result.

Table 11: Confusion matrix for subclustering doctors and nurses

	Cluster 1	Cluster 2
Doctor	14	0
Nurse	6	1

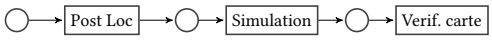
(a) Activity-based clustering

	Cluster 1	Cluster 2
Doctor	14	0
Nurse	1	6

(b) LPM clustering

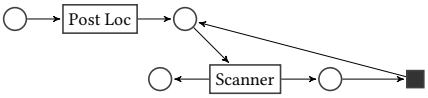
Figures 3 and 4 show four relevant LPMs coming from the dataset. They highlight the difference in behavior between the nurses and the doctors. LPMs 3a and 4a are performed by nurses whereas LPMs 3b and 4b are performed by doctors. The same activities are performed in both cases but in different ways. Qualitatively, the LPMs provide a better understanding of the processes.

(a) Nurses

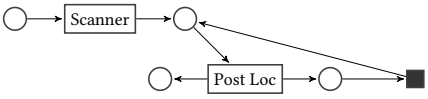


(b) Doctors

Figure 3: Simple LPMs extracted from the real-life dataset



(a) Nurses



(b) Doctors

Figure 4: Complex LPMs extracted from the real-life dataset

5 LIMITATIONS AND FUTURE WORK

5.1 Variation of Distribution

Based on the synthetic dataset, we will evaluate the impact on the results of a more unfair distribution while keeping a fair distribution of the workload. Two cases will be considered:

- The over-representation of one class.
- The under-representation of one class.

The clustering based on the activities is completely quite insensitive to the variation of distribution. This clustering does not consider the links between the activities and then cannot understand the difference between the classes N1, N3 and N4. The case exposed in Subsection 3.4 is consequently already a case of unfair distribution.

The clustering based on the LPMs is quite insensitive to the variation of distribution (see also next subsection). Its sensitivity comes from the LPM extraction which can be impacted by the distribution variation. Indeed, the frequency of appearance is taken into account to choose the most relevant LPMs and then an unfair distribution of the instances will lead to an unfair distribution of the LPMs. This will decrease the clustering quality as exposed in Figure 5a where α is the relative size of the instance set of N2 with respect to the instance set of any other class.

The solution used to improve the results is to increase the number of extracted LPMs as depicted in Figure 5b. There must be more extracted LPMs than the LPMs required to model the behavior of the over-represented class to represent optimally the situation. There are two drawbacks: the distribution is a priori unknown and a too high increase of the number of LPMs will lead to overfitting.

The under-representation a of class brings exactly the same issues with an extreme case where none of the extracted LPMs is coming from this under-represented class. Increasing the number of extracted LPMs will not always solve this because of the under-representation of only one class. Indeed the additional LPMs could still avoid the under-represented class.

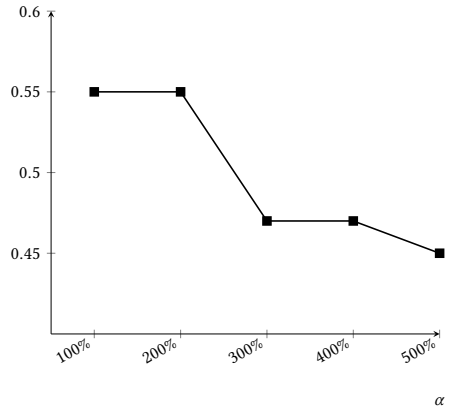
5.2 Robustness to Outliers

This subsection will discuss the impact of the presence of outliers in the dataset. Two kinds of outliers are modeled:

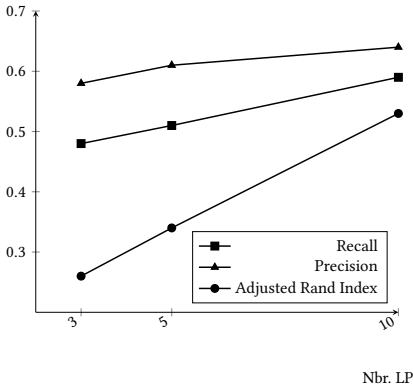
- Irrelevant executions of the process resulting from errors in the encoding (duplicate activities, missing activities, etc.). Those errors are encountered frequently in real-life process.
- Execution of the process in an unexpected way. The process is consistent but is not coming from the synthetic dataset. For example, it can be nurses with a hybrid behavior, working sequentially on a patient then deciding to work in batch on the two next patients.

Based on the knowledge of the ground-truth, we will add a new cluster. The purpose of this approach is to show that the outliers are correctly identified and classified in this new cluster. Once again, only the nurses will be considered.

5.2.1 Irrelevant Executions. Those executions are based on encoding errors leading to several missing or falsely repeated activities. The approach using the activities is based on their frequencies and then is directly impacted by errors in the encoding. The LPM-based



(a) Recall regarding to the distribution



(b) Recall and precision for different numbers of extracted LPMs

Figure 5: The influence of the distribution on clustering performances

Table 12: Outliers

	Trace
Normal	**A** - B - C - **A** - B - C - **A** - B - C
Irrelevant	**A** - B - C - **A** - A - B - C - **A** - B - C
Unexpected	**A** - B - C - **A** - **A** - B - **B** - **C** - **C**

approach takes into account the sequencing of the activities, it extracts recurring patterns which are not directly affected by irrelevant activities to some extent. The proposed approach is more robust regarding irrelevant executions.

5.2.2 Unexpected Executions. The approach based on the activities is unable to identify coherent unexpected executions because this approach does not consider the links between the activities and then the coherence of the execution. The only difference between a common execution and an unexpected execution is how the activities are performed.

By considering patterns rather than activities, the proposed approach, as depicted in Table 13, identifies the outliers. The limitation of the approach is that outliers are an extreme case of unfair distribution and the issue discussed in the previous subsection are also applicable in this case.

Table 13: Outliers comparison

Frequency of:	A	B	C
Normal	3	3	3
Irrelevant	4	3	3
Unexpected	3	3	3

(a) Activity clustering

Frequency of:	A B C	A A	B B	C C
Normal	3	0	0	0
Irrelevant	3	0	0	0
Unexpected	1	1	1	1

(b) LPM clustering

5.3 Workload

Throughout this paper, the assumptions of a perfectly fair distribution of the workload and of a clearly defined start point/endpoint were made. This is not necessarily the case in real datasets. Each resource may have its own working time. Even if they have the same behavior, two resources will not perform the same number of activities/tasks.

Future work aims to evaluate the results of the LPM-based approach on resources represented by longer traces and with different time horizons.

6 CONCLUSION

The purpose of this paper was to present an approach allowing us to identify people working in different ways, to detect their behaviors and to cluster them based on these behaviors. Based on a realistic healthcare example, this paper introduced the key concepts of our approach. The traces representing the activities performed by the workers and the local process models providing a modeling of the behaviors.

The results of our approach, compared to the results of the existing approaches based only on the activities, show that LPMs improve significantly the quality of the clustering. For trivial situations where the resources perform different activities (Doctor vs Nurse), both approaches provide the same good results but only our approach is able to distinguish the resources performing the same activity in different ways (sequential vs in batch vs as they come). The quality of the clustering is even more increased when the notion of case (the patient) is incorporated.

We highlight that the extraction of LPMs and by extension the approach is sensitive to the distribution of the kinds of resources. The extraction of the LPMs is a key point of the approach and is significantly impacted by a range of factors such as distribution, workload and outliers. These factors will be investigated in future researches.

Applying our proposal to an authentic datasets coming from the hospital field confirms that the LPMs improve the quality of the clustering by identifying resources performing the same activities but in a different way.

REFERENCES

[1] Camilo Alvarez, Eric Rojas, Michael Arias, Jorge Munoz-Gama, Marcos Sepúlveda, Valeria Herskovic, and Daniel Capurro. 2018. Discovering role interaction models in the Emergency Room using Process Mining. *Journal of biomedical informatics* 78 (2018), 60–77.

[2] Ana Karla Alves De Medeiros, Antonella Guzzo, Gianluigi Greco, Wil Van Der Aalst, AJMM Weijters, Boudewijn F Van Dongen, and Domenico Saccà. 2007. Process mining based on clustering: A quest for precision. In *International Conference on Business Process Management*. Springer, 17–29.

[3] El Mostafa Qannari, Philippe Courcoux, and Pauline Faye. 2014. Significance test of the adjusted Rand index. Application to the free sorting task. *Food quality and preference* 32 (2014), 93–97.

[4] Stefan Schönig, Cristina Cabanillas Macias, Claudio Di Ciccio, Stefan Jablonski, and Jan Mendling. 2016. Mining resource assignments and teamwork compositions from process logs. *Softwaretechnik-Trends* 36, 4 (2016), 1–6.

[5] Minseok Song, Christian W Günther, and Wil Van der Aalst. 2008. Trace clustering in process mining. In *International Conference on Business Process Management*. Springer, 109–120.

[6] Minseok Song and Wil Van der Aalst. 2008. Towards comprehensive support for organizational mining. *Decision Support Systems* 46, 1 (2008), 300–317.

[7] Niek Tax, Natalia Sidorova, Reinder Haakma, and Wil van der Aalst. 2016. Mining local process models. *Journal of Innovation in Digital Ecosystems* 3, 2 (2016), 183–196.

[8] Wil Van Der Aalst. 2011. *Process mining: discovery, conformance and enhancement of business processes*. Vol. 2. Springer.

[9] Wil Van Der Aalst, Hajo A Reijers, and Minseok Song. 2005. Discovering social networks from event logs. *Computer Supported Cooperative Work (CSCW)* 14, 6 (2005), 549–593.

[10] Wil Van der Aalst, Ton Weijters, and Laura Maruster. 2004. Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge & Data Engineering* 9 (2004), 1128–1142.