PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Lucas Seixas Jazbik

Deviance mining of online processes with non-atomic events in the COVID-19 domain

Dissertation presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção

Advisor: Prof. Fernanda Araujo Baião Amorim

Rio de Janeiro,

September 2022

PUC-Rio - Certificação Digital Nº 2020862/CA

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Lucas Seixas Jazbik

PUC-Rio - Certificação Digital Nº 2020862/CA

Deviance mining of online processes with non-atomic events in the COVID-19 domain

Dissertation presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção. Approved by the Examination Committee:

Prof. Fernanda Araujo Baião Amorim
Departamento de Engenharia Industrial - PUC-Rio

Prof. Hélio Côrtes Vieira Lopes
Departamento de Informática - PUC-Rio

Prof. Luiz Felipe Roris Rodriguez Scavarda do Carmo
Departamento de Engenharia Industrial - PUC-Rio

Rio de Janeiro, September 1, 2022

Lucas Seixas Jazbik

The author graduated in Production Engineering from Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio in 2019, co-authored two scientific articles submitted to highly-qualified international Engineering journals (one of them already published) and works at Ambev since 2019.

PUC-Rio - Certificação Digital Nº 2020862/CA

Ficha Catalográfica

Jazbik, Lucas Seixas

Deviance mining of online processes with non-atomic events in the COVID-19 domain / Lucas Seixas Jazbik ; advisor: Fernanda Araujo Baião Amorim. – 2022.

69 f. : il. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2022.

Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Processos Contínuos. 3. Eventos não atômicos 4. Melhoria de processos. 5. COVID-19. I. Baião Amorim, Fernanda Araujo. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. III. Título.

CDD: 658.5

PUC-Rio - Certificação Digital Nº 2020862/CA

Acknowledgements

First and foremost, I would like to praise and thank God, because without him nothing is possible.

Apart from the efforts of me, the success of this thesis depends largely on many others. I take this opportunity to express my gratitude:

To my adviser, Dr. Fernanda Araujo Baião, for her patience and encouragement throughout this journey.

To PUC-Rio for the support granted, without which my graduation and master's degree would not have been possible.

To my parents, family and friends for all their support.

To Ana Clara Quinhões for all her love.

To the teachers of the examining board and to all the teachers who were part of my education.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Jazbik, Lucas Seixas; Baião, Fernanda. **Deviance mining of online processes with non-atomic events in the COVID-19 domain** . Rio de Janeiro, 2022. 69 p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Process Mining techniques have been successfully applied as a data-driven and domain-aware approach for improving business process performance in several organizations. Among its applications, Deviance Mining aims at uncovering the reasons why a subset of the executions of a business process deviate with respect to its expected or desirable outcomes, thus producing insights towards improving the process operation, such discoveries can be made using treatment learning techniques, which identify the sets of attributes that are most influential in the results. However, despite the fact that real-life processes are typically composed by events with non-instantaneous duration (non-atomic events), existing approaches for process mining and deviance mining in particular only deal with atomic events in their experiments. This work proposes a domain-driven method for automatically detecting deviations in processes composed by non-atomic events. The method uses the temporal dimension of non-atomic events to apply deviance mining, generating insights on how the duration and the simultaneous occurrence of events generate deviations and how these deviations affect the results of the processes. The method was successfully applied in the COVID-19 domain, to find which domain-specific sequences of non-pharmaceutical interventions mostly contributed to slowing down the rate of COVID-19 cases in countries around the world.

PUC-Rio - Certificação Digital Nº 2020862/CA

Keywords

Continuous Process, Non-atomic events, Process improvement, Deviance mining, COVID-19

PUC-Rio - Certificação Digital Nº 2020862/CA

Resumo

Jazbik, Lucas Seixas; Baião, Fernanda. **Mineração de desvios em processos online com eventos não atômicos no domínio da COVID-19**. Rio de Janeiro, 2022. 69 p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

As técnicas de mineração de processos vêm sendo aplicadas com sucesso como abordagens baseadas em dados e específicas do domínio para melhorar o desempenho do processo de negócios em várias organizações. Dentre suas aplicações, a Mineração de Desvios (*Deviance Mining*) visa descobrir as razões pelas quais um subconjunto das execuções de um processo de negócio desvia-se em relação aos seus resultados esperados ou desejáveis, produzindo assim *insights* para melhorar a operação do processo, tais descobertas podem ser feitas utilizando técnicas de aprendizado de tratamentos (*Treatment Learning*), que identificam os conjuntos de atributos mais influentes nos resultados. No entanto, apesar de os processos da vida real serem tipicamente compostos por eventos de duração não instantânea (eventos não atômicos), as abordagens existentes para mineração de processos, e para mineração de desvios em particular, endereçam exclusivamente eventos atômicos em seus experimentos. Este trabalho propõe um método orientado ao domínio para detectar automaticamente desvios em processos compostos por eventos não atômicos. O método utiliza a dimensão temporal dos eventos não atômicos para aplicar a mineração de desvios, gerando insights sobre como a duração e a ocorrência simultânea de eventos geram desvios e como esses desvios impactam os resultados dos processos. O método foi aplicado com sucesso no domínio da COVID-19, para descobrir quais

PUC-Rio - Certificação Digital Nº 2020862/CA

sequências de intervenções não farmacêuticas mais contribuíram para diminuir a taxa de casos de COVID-19 em países ao redor do mundo.

Palavras-Chave

Processos Contínuos, Eventos não atômicos, Melhoria de processos, Mineração de desvios, COVID-19

PUC-Rio - Certificação Digital Nº 2020862/CA

Sumário

PUC-Rio - Certificação Digital Nº 2020862/CA

Lista de figuras

PUC-Rio - Certificação Digital Nº 2020862/CA

Lista de tabelas

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

1 Introduction

Process mining uses data from the historical executions of a process, recorded in an event log, to extract process models denoting execution patterns about the studied process. Existing process mining techniques are not based on abstract ideas of how a process should happen, but on real execution data that represent how the process actually occurs [35].

Fig. 1 shows an illustrative example using the Business Process Modeling Notation (BPMN) [52] of a process of a patient's visit to a hospital. This model is not able to provide various data about the represented process. For example, it is not known whether all the flows that a patient may follow are represented in the model, if there is flexibility in the execution of the process in relation to the model, or if this model precisely fits the actual process executions. To formulate knowledge that genuinely reflects the actual executions of a process, process mining techniques are applied [36].

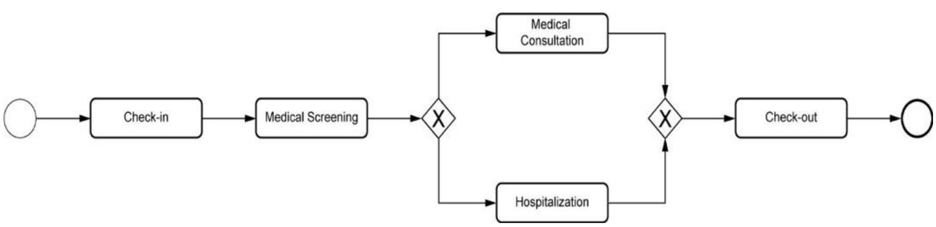


Fig. 1 Example of a process model of a hospital visit, in BPMN notation

Among the process mining sub-areas, deviance mining aims to identify and analyze deviations in the execution of a process in relation to expected behavior. Such deviations can generate effects considered to be positive or negative to the outcome of the analyzed process [28]. The choice of a performance indicator of the analyzed process (e.g., execution time, total cost, etc.) as the class variable determines which process output will characterize the deviations. The use of treatment learning techniques [25, 26] helps in the characterization of deviations in relation to the class variable, as it encompasses techniques that allow identifying sets of process attributes which most influence the results. The present work proposes a new method that uses treatment learning to identify recurring consequences of process deviations.

It is well known that the great majority of real processes are constituted by non-instantaneous events, which are called non-atomic events (i.e., events that have non-instantaneous duration) [1]. Despite the predominance of non-atomic events in real life, there are few works in the process mining literature regarding applications to processes with non-atomic events [2]. The proposed method uses the temporal dimension of non-atomic events to generate insights on how the duration and the simultaneous occurrence of events generate deviations and how these deviations affect the results of the processes. By addressing the non-atomic characteristic of the processes, this work brings the application of process mining closer to scenarios that are semantically more precise with respect to reality.

The process represented in Fig. 1 is a classic example of a process mining application with well-defined beginning and end. However, there are situations where the application of process mining occurs continuously (i.e., never ending), in what are called online scenarios. Because it is an endless process or a process whose executions are still in progress [7, 23, 38], such situations increase the complexity of the studied processes for the purpose of identifying deviations. Our proposed method allows for the application of deviance mining in an online scenario, so it differs from traditional deviance mining applications [28].

PUC-Rio - Certificação Digital Nº 2020862/CA

Due to the large number of variables present for analysis in this work, as a way to increase the effectiveness of the results, a feature selection step was required in the proposed method. The feature selection problem consists of selecting the attributes relevant to the solution of the problem, to maximize the value of the results [17]. Such problems are very costly to solve by an exhaustive search, so there are several metaheuristic applications for their solution [41]. For the feature selection step of the proposed method, a new metaheuristic designed for the context of this work was created and its effectiveness was tested.

To validate the method proposed in this work, we used a process composed of atomic and non-atomic events, from an online scenario. The chosen scenario was the COVID-19 pandemic, where non-pharmaceutical containment measures taken by several countries can be understood as process events, whose result is their impact on the acceleration of the number of individuals infected by the virus (i.e., COVID-19 cases). To the best of our knowledge, no other studies have approached the global fight against the pandemic as a process using process mining to extract knowledge from it.

The proposed method was evaluated concerning the quality and interpretability of the discovered treatments, and concerning its computational performance (when comparing its application with and without the proposed feature selection metaheuristic). The evaluation of the method from different perspectives aimed to prove the effectiveness, efficiency and applicability of the proposal.

The fight against COVID-19 by different countries around the world was selected as the application domain for evaluating the proposed method since it is a continuous, complex, and flexible process, with the collaboration of several actors and events, so that the effectiveness of the proposed solutions could be tested in a real, knowledge-intensive

PUC-Rio - Certificação Digital Nº 2020862/CA

scenario. However, the method proposed in the present research, as well as the computational tools that were developed to support its execution, are not limited to an exclusive application to this domain and can be applied in several processes. In this way, we argue that this work contributes to Industrial Engineering by facilitating decision makers who act as process managers, whether of greater or lesser complexity, including, for example, supply chain management processes and organizational management processes.

The main contribution of this work is the proposal of a method that, through its pre-processing stage, its feature selection metaheuristic, and the application of treatment learning techniques, automatically discovers intelligible rules that enhance the best results of the analyzed process (or avoid its worst results), thus supporting the process managers' decision on how to improve the process to achieve better performance indicators. The proposed method considers the temporal dimension of the events that compose this process and the frequent changes that may be observed among its instances.

This work is structured as follows. Section 2 presents the definitions and techniques we apply, as background knowledge for understanding our proposed approach. Section 3 discusses existing works in the literature related to our proposal, which is detailed in Section 4. Experimental results are presented in Section 5, and Section 6 summarizes our main conclusions and recommendations for future research.

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

2 Background Knowledge

This section presents fundamental concepts about online process mining, atomic and non-atomic events, deviance mining, and treatment learning, which are required for a good comprehension of the proposed method.

2.1 Process mining definitions

According to van der Aalst et al. [36], process mining is a research topic involving data mining and process modeling that helps organizations to find and compare models, and develop better processes. Process mining uses event logs as inputs to extract insights in the form of models, frequently control-flow models. Event logs record the actual behavior of a process, in the form of process traces composed of the sequence of events that were executed during a process. Definitions 1 to 3 formally specify event logs and their composing elements.

Definition 1 (Process Event): A process event e represents the occurrence of an activity during the execution of a process. An event is characterized by the tuple e = (eid, eattr), where eid is the unique identifier of the occurrence of event e and eattr is a set of event-level attribute-value pairs ((a1,v1),...,(aj, vj)), where j > 0.

Definition 2 (Process Trace): A process trace t (a.k.a. "process instance" or "case") is characterized by the tuple $t = (t_{id}, t_{attr}, \sigma)$, where t_{id} is the unique identifier of the trace, t_{attr} is a set of trace-level attribute-value pairs $((a_1,v_1),...,(a_m, v_m))$ where $m > 0$, and $\sigma = \{e_1, e_2, ..., e_n\}$ is a finite sequence of events e_i that occurred during the process instance execution. Note that σ must have at least one event ($|\sigma| > 0$).

Definition 3 (Event Log): Given a single business process *P*, an event log *L* is characterized by the tuple $L = (T, P)$, where *T* is a set of process traces $\{t_1,..., t_n\}$ left by executions of process *P* and $t_i \in T$, $1 < i < n$.

2.2 (Non-)atomic events

As formally specified in Definition 1, the attributes (and associated values) of an event represent the relevant characteristics of its execution, which may be related to the cost, the person who executed it, or any other event aspect that was recorded in the log. A common category of event attributes is related to the temporal dimension [34]. Such attributes can simply be time stamps that indicate the beginning of some event, or the duration of an event. An event whose duration is zero or non-existent is called an atomic event (i.e., an instantaneous event).

According to Bernardi et al. [1], most process mining approaches simplify event characterizations by assuming all events of a process are atomic, even though the vast majority of real-life processes are composed of non-atomic events. Non-atomic events are not instantaneous, since they occur over a period of time and have a life cycle [2]. When considering the existence of the event life cycle (sequence of transactional states assumed by a non-atomic event) and its duration, we approximate the studied scenarios to their real characteristics. Therefore, we refine Definition 1 (which still holds specifically for atomic events) into Definition 4.

Definition 4 (Non-atomic Event): A process's non-atomic event *nae* represents the non-instantaneous occurrence of an activity during the execution of the process. A non-atomic event is characterized by the tuple $nae = (nae_{id}, nae_{attr}, nae_{states})$, where nae_{id} is the unique identifier of the occurrence of event *nae*; nae_{attr} is a set of event-level attribute-value pairs $((a_1,v_1),...,(a_j, v_j))$ with $j > 0$

PUC-Rio - Certificação Digital Nº 2020862/CA

where one of the attribute-value pairs *(a_i, Δ)* denotes the execution time with $\Delta > 0$; and *nae_{states}* is a sequence of transactional states *<nae start,...,nae finish>* assumed by event *nae* during its execution.

The challenge for process mining techniques to address non-atomic events is to handle not only the duration aspect of an event, but also the simultaneous occurrence among them, since at the same time instant there may be several events occurring in parallel, and either elapse of time or the co-occurrence of events may impact the process outcome. As presented in [2, 40], the study of parallelism between event occurrences can lead to several process improvements, and both time lapse and parallelism in execution are characteristics that make the studied scenarios semantically more precise with respect to their real-world counterparts. However, according to Cheng et al. [8], process mining tools present difficulties in working with these issues.

2.3 Online process mining

Process mining can be characterized in two ways with respect to the continuity of the process being mined, which is reflected in the structure of the input data: offline process mining and online process mining. Offline process mining uses a static event log as input, thus using data from already completed process instances [38]. Online process mining, on the other hand, analyzes the execution of an unfinished process, thus taking place during the execution of the process, using the data recorded in a continuous data stream (see Definition 5) to extract knowledge about the process [7].

PUC-Rio - Certificação Digital Nº 2020862/CA

Definition 5 (Data Stream): Given a single real-time business process *P*, a data stream *DS* is characterized by the ordered pair $DS = (S, \Delta)$, where *S* is a sequence of events of process P and Δ is a sequence of positive real-time intervals that indicate the time when each event occurred.

Similarly to its offline counterpart, online process mining can also be applied to conformance checking, process discovery and process improvement. One of the distinctions between the two areas is that a process which occurs in an online scenario denotes either a process that has not yet ended, or a never-ending one (i.e., a continuous process such as ”Data Quality Control”, or ”Monitoring a Pandemic”). This requires a different approach from the one applied for offline process mining. The main difference between the two approaches is that for offline process mining, the instances of a process have already been completed, facilitating the identification of the outcome associated with each process instance, while in the case of online process mining, the instances are continuously running, making it difficult to identify the influence of the process characteristics on the execution, and the outcome of the process itself.

An important issue for online process mining is thus to determine the analytic intervals, that is, the periods of time during which the events of a process constitute a complete instance. To this end, our proposal defines analysis intervals by detecting concept changes in the process. According to Martjushev et al. [24], processes can be modified according to three main perspectives: control-flow, data and resources. An analysis interval is defined when a significant change in one of these perspectives of the process is identified. When this change is identified, the corresponding event is called a change point. According to Bose and van der Aalst [3], three steps are required to address these significant changes:

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

- **Change point detection:** The first and most fundamental step is to detect the changes that have occurred and when they occurred.
- **Change localization and characterization:** The second step is to characterize the changes that have taken place and identify their locations in the process. This step is a challenge that involves discovering the pace of change (sudden, gradual, etc.) and identifying the change itself within the process (the instances, dates and specific events that occurred).
- **Change process discovery:** After identifying, locating and characterizing the change, the third step puts it in perspective. At this point, the tools are applied to explore and relate these discoveries. Understanding the evolution of a process should lead to understanding the process of change, by describing the second-order dynamics of the process.

The identification of change points enables the definition of analysis intervals of a process within an online scenario, thus defining the traces of this process to be mined and making it possible to analyze the outcome of each trace. The analysis intervals defined for an online scenario are named *windows of analysis* [7]. To improve the accuracy of the change point definition, statistical tests are used to define the windows of analysis [20, 21].

Definition 6 (Window of Analysis): A window of analysis W can be defined using its start time Ws and its end time We. In comparing two windows Wa and Wb, we can say that Wa precedes Wb, formally $Wa<Wb$, if $Wae<Wbe$. L^W denotes the projection of an event log L to a window W.

2.4 Deviance mining and treatment learning

The analysis of deviations along process executions can generate insights to their executors regarding the characterization and effects of these deviations [34]. Such analysis is carried out from a performance perspective, hence the deviations are characterized in relation to their impact on the process outcome. In this way, deviance mining is the set of techniques to identify and characterize deviations in the execution of processes according to the consequences (either positive or negative) of such deviations on the predicted outcome. Such techniques aim to identify events that lead to deviations from the analyzed process, generating information that leads to improvements in the model and in the execution of the process [28].

Traditionally, deviance mining is conducted using unsupervised or supervised strategies. In the first case, there is no prior information regarding cases considered to be deviations. Moreover, unsupervised strategies are divided into two main groups: "model-based" and "clustering or similarity based". Unsupervised model-based approaches use conformance checking techniques [11] to compare the execution of a process with its reference model, in order to classify the executions as normal or deviations. Similarity and clustering-based approaches do not use reference models and assume that normal traces occur in large clusters and that deviations occur in small or no clusters [12]. This approach identifies deviations by comparing each trace with its neighbors through similarity, distance or density functions [19]. In supervised deviance mining, all traces are characterized according to some punctuation or deviation marking, in order to clearly define which cases are process deviations. Techniques such as decision trees may then be applied [16],

PUC-Rio - Certificação Digital Nº 2020862/CA

but their results can be extremely complex and hard to interpret. To improve the interpretation of the deviance mining results, Richetti et al. [30] proposed an approach for supervised deviance mining that applies a treatment learning algorithm to discover rules (named treatments) that represent deviations, by analyzing their effects on the distribution of the class variables that represent the outcome of the analyzed process. In treatment learning, treatments are the smallest possible attribute baskets that differentiate instances of a dataset associated with a higher-weight class from instances associated with a lower-weight class [25].

Definition 7 (Dataset): A dataset D is a set of n data instances, $D=\{i_1,...,i_n\}$, $n>0$. Each data instance i is characterized by the tuple $i = (i_{id}, i_{attr}, c_{name})$, where i_{id} is the unique identifier of a data instance, i_{attr} is a set of attribute-value pairs $((a_1,v_1),...,(a_k, v_k))$, where $k > 0$ is the total number of attributes in D, and c_{name} is the name of the class attribute, corresponding to the name of a single attribute $a' \in (a_1 ... a_k)$.

Association-rule approaches are used for treatment learning [16, 26] and have some advantages over other classification techniques, since they do not assume a specific distribution of data, tend to generate intelligible rules (treatments) and are able to work with unbalanced datasets in relation to the [3] class variable. Many association-rule techniques adopt a support-based criterion to prune less frequent treatments. This strategy can lead to the exclusion of relevant treatments and the discovery of only obvious treatments. According to Hu [16], the use of a criterion based on trust can lead to interesting discoveries.

Given a dataset D, a treatment learning tool looks for a treatment R_X that returns a subset of attributes $D' \subseteq D$, where $D' = \{D \cap R_X\}$, so that the frequency of the preferred class is maximized relative to the original distribution of D [16].

PUC-Rio - Certificação Digital Nº 2020862/CA

Consider the illustrative dataset from Table 1, which represents the instances of a borrowing process. The purpose of the analysis is to discover which dataset attributes values implying the highest cost per loan. Therefore, the attribute ”cost per loan” was defined as a class variable for this process. Suppose the distribution of the illustrative dataset class variable follows the baseline distribution shown in Fig. 2 (left). A treatment learning tool could discover the R_1 *treatment:* {*market sector = ”energy”*}. In a *D’* subset of the original dataset where this treatment holds, there is a change in the class variable distribution, as shown in Fig. 2 (right).

Table 1 Example of a dataset for treatment learning.

Case Instance	*Principal*	...	*Market Sector*	*Cost per Loan*
#1000	\$50K	...	Real estate	Low
#1001	\$10K	...	Real estate	Medium
#1002	\$100K	...	Consumer staples	Low
#1003	\$1M	...	Energy	High
...	...	...	...	...

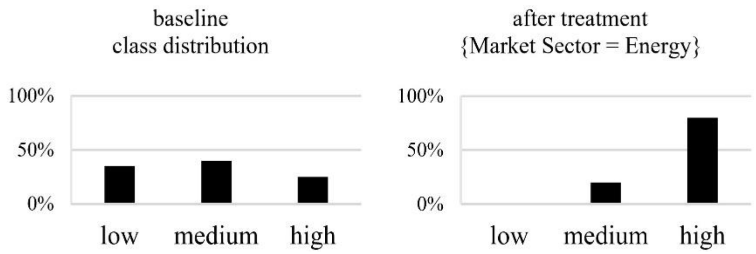


Fig. 2 Baseline distribution (left) and after treatment (right) [30].

PUC-Rio - Certificação Digital Nº 2020862/CA

2.5 On Process Mining and interpretability in Knowledge-intensive scenarios

The recent and increasingly frequent advances in the application of Process (and Data) Mining techniques arose the issue of explainability of the results. Many Data Science tools face reluctance to be applied in organizations because decision makers, who may use their results to guide their decisions, do not easily understand these results or their functioning [48].

Interpretability of the methods and results obtained through Artificial Intelligence and Data Science tools is a concern addressed by several studies on "Explainable AI" and applied to several domains, including Medicine and Engineering [45, 46, 47, 48]. In order to increase the chances of success in the application of new technologies, it is necessary that organizations overcome the resistance of users to change and present methods and tools that are traceable and whose behavior is understandable [49].

In order to facilitate the understanding of complex or poorly structured processes in process mining, declarative process modeling emerged, in particular using the declare language [50]. This constraint-based language produces process models represented as a set of rules, that is, in which only mandatory flows and prohibitions are explicitly mapped, leaving all other possible execution flows open to the user. This has been proven as a successful approach to avoid the creation of extremely complex diagrams showing all possible execution flows (a.k.a. spaghetti diagrams) [51] .

PUC-Rio - Certificação Digital Nº 2020862/CA

The use of Treatment Learning techniques, as proposed in this work, is also a way to automatically discover correct and easier to interpret results, when compared with the use of other methods such as Random Forests (since these methods also tend to generate spaghetti diagrams when being applied on complex processes [30]).

Thus, data science methods that seek to achieve success in organizations must offer tools and results that present good performance, are prepared to work with large amounts of data (scalability) and present understandable and explainable results that guide actions to be taken [48].

PUC-Rio - Certificação Digital Nº 2020862/CA

3 Related Work

This section discusses existing works related to online process mining and compares the method proposed in this work with other approaches in the literature.

Bozorgi et al. [4] proposed a prescriptive monitoring method to decide in which cases to intervene during the execution of a process so as to improve its overall net gain. The method is divided into an offline phase (to discover a set of interventions that generally improve the process, using historical data) and an online phase (which monitors the process execution and applies user-defined policies to trigger recommendations on when to apply a previously discovered intervention in an ongoing case). There are similarities between the referred approach and our proposal, since both target the discovery of improvements to the process from historical cases (interventions, or in our case treatments), but an important distinction is that their proposal does not address non-atomic events or the need to define traces from a continuous flow of events generated by the process, which are two issues specifically addressed by our proposal.

The works of Lee et al. [18] and Zaman et al. [42] proposed frameworks for conformance checking of processes in online scenarios. Both works concern memory constraints for data storage (which is not the focus of our approach), which is a common, yet complementary, issue in online scenarios, which typically deal with data streams and real-time data [15, 31].

The work of Ceravolo et al. [7] identified a set of goals for online process mining approaches, and indicated the concept drift detection task (and in particular the definition of analysis windows) as a critical issue

PUC-Rio - Certificação Digital Nº 2020862/CA

for online process mining. The authors did not consider the presence of non-atomic events in the process log. Our proposal, in turn, specifically addresses the definition of analysis windows and also considers the existence of both atomic and non-atomic events.

Burattin et al. [5] and Navarin et al. [27] presented approaches for real-time model discovery using the Declare notation, so that process analysts can access process information instantly and in a language that facilitates the interpretation of the model. In particular, Navarin et al. [27] contributed by incorporating both control flow dependencies and data conditions to discover declarative process models from event streams, while Burattin et al. [5] focused on the discovery of LTL-based constraints. Our approach, on the other hand, proposes and evaluates the codification of features of a specific nature in the dataset (the duration and simultaneous execution of non-atomic events), while keeping the possibility of including additional domain-specific features, such as those described in [27], [5] and [30].

Similarly to offline scenarios, concept drifts also impact mining processes occurring in online scenarios, since changes in the process can make all analyses and discoveries outdated. Thus, several techniques exist to deal with concept drifts in online scenarios [6]. In particular, some process improvement approaches address concept drifts, such as Spenrath and Hassani [32], who proposed a model to predict bottlenecks in the execution of a process in online data streams that is adapted to the occurrence of concept drifts, so that predictions remain correct. Hassani [14] presented an approach for defining the windows of analysis and their sizes according to the characteristics of the concept drift, so that the data to be analyzed actually represent the process. Maaradji et al. [22] proposed an efficient method to detect sudden and gradual concept drifts of process executions. Ostovar et al. [29] presented a method to define

PUC-Rio - Certificação Digital Nº 2020862/CA

analysis windows in processes with a high degree of variation in their executions. Our proposed method is also able to deal with concept drift situations, since it uses the concept of analysis windows to define how to partition data.

Similar to our proposed method, Montali et al. [55] presents a framework for considering the temporal dimension of the process events in online scenarios. This framework, named Mobucon EC, is based on streams of events to dynamically monitor the compliance of running cases with the business rules of complex processes. Mobucon EC uses a time-aware extension of the declare language to address constraints regarding the duration of process activities. The authors highlight the importance of observing the process in real time and considering its temporal dimensions.

PUC-Rio - Certificação Digital Nº 2020862/CA

4 Proposed Method

This section presents the proposed method for deviance mining of online processes with non-atomic events through treatment learning. Fig. 3 represents the method's steps with associated input and output resources. The method starts by detecting change points in the event log of the online process, in order to define windows of analysis (as explained in Section 2.3), which correspond to the process traces to be analyzed. With the process traces defined, the event log of the process needs to be converted into a table-like dataset for treatment learning. For this conversion, the proposed method requires three steps: first, a feature engineering step creates features corresponding to potentially relevant characteristics of the process, including: (i) the class variable represented by the process outcome at the time of the change point detected; (ii) a set of event attributes from the event log, in particular the duration of each non-atomic event's occurrence; and (iii) a set of attributes representing the parallelism between each pair of event occurrences. Since the number of generated features can be extremely large and therefore prevent the execution of the deviance mining algorithm in a reasonable computing time to obtain relevant results, the second step consists of a metaheuristic algorithm to perform feature selection so as to select the most relevant subset of features from the dataset that will improve the effectiveness of the treatment learning, which is the final step of the method, responsible for discovering treatments that correspond to the process deviations. The method's steps are further detailed in the following subsections.

PUC-Rio - Certificação Digital Nº 2020862/CA

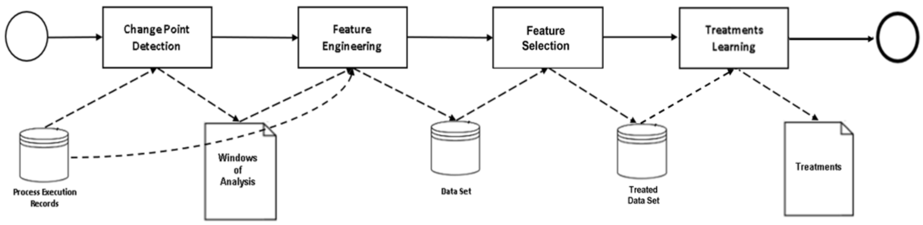


Fig. 3 Flow of activities composing the proposed method, illustrated as a BPMN process

4.1 Change point detection

Online process mining requires the definition of windows of analysis, in order to represent time ranges within which the process evolution occurred and enable an analysis of how this behavior may have impacted the outcome of the process, measured at the end of each analysis period. Therefore, the windows of analysis characterize the process traces to be mined.

To define the windows of analysis of an online process, we assume the event log contains one attribute representing the process outcome, whose value is periodically registered over time. Typically, this attribute constitutes a performance indicator of the process [9, 37] that was defined by the process manager as the focus of the process analysis and of the deviation detection. This chosen attribute will constitute the class variable of the dataset, and it is possible to trace the time series of the chosen class variable to detect its change points, or deviations.

Our proposed deviance mining method seeks to understand the effect of the identified deviations on the result of the process. Thus, the definition of the windows of analysis must occur from the changes identified in the time series of the class variable, so that each window of analysis will contain the events that occurred between a first change point, which marks the beginning of the window of analysis, and a second

PUC-Rio - Certificação Digital Nº 2020862/CA

change point, which marks its end. We state that the change points in the time series of the class variable are the points where a structural change [33] in the time series occurs. The identification of change points occurs through statistical tests (fluctuation tests and F-tests), present in the R system [1]**strucchange** package [43].

The windows of analysis are defined as a product of this step. These windows will represent the dataset instances. In order to understand how the events of a process instance affect the result of the process, each window of analysis is associated with the result that the process presented at the time of its occurrence. For this reason, each analysis window is associated with the value that the class variable had at the change point representing the end of that window.

PUC-Rio - Certificação Digital Nº 2020862/CA

4.2 Feature engineering

After defining the windows of analysis, which form the instances of the dataset developed, it is necessary to define the attributes that will form this dataset. For treatment learning to discover treatments that generate significant insights in relation to the studied process, it is necessary for the attributes that will form the dataset to be representative in relation to the characteristics of the process they are intended to represent. When dealing with processes with non-atomic events, the temporal dimension is extremely significant in relation to the studied process [2], and should be represented in the dataset.

The execution time of the events of a process, as well as the simultaneous occurrence of these events within a trace, are important characteristics to understand and represent processes with non-atomic events [2, 40], which are potentially relevant for analysis. The execution

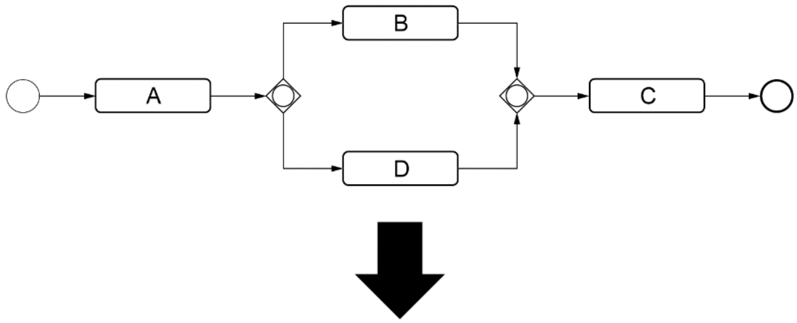
[1] https://www.r-project.org/

time of an event allows us to discover how the duration of multiple events in a process impact the outcome of that process. In turn, the simultaneous occurrence of events in a trace allows us to discover how pairs of events occurring in parallel can affect the outcome of that process. Thus, to better represent processes with non-atomic events for the purpose of deviance mining, it is imperative for these two variables to be codified as attributes in the dataset.

Fig. 4 illustrates a fictitious process with non-atomic events (composed by the events ”A”, ”B”, ”C” and ”D”, where events ”B” and ”D” can occur simultaneously) and a small example of the dataset codified for some of the process traces, after carrying out steps 1 (”Change Point Detection”) and 2 (”Feature Engineering”) of the proposed method. Each event (”A”, ”B”, ”C” and ”D”) is codified into an attribute representing its duration within the trace. The value of the process indicator when the analysis window ends is also codified as an attribute (the class variable). Additional attributes are inserted in the dataset to represent how long a pair of events occurred simultaneously (such as the attribute ”Simultaneous duration of events B and D” - which may occur in parallel according to the process model). Finally, if there are additional characteristics in the process that are judged by the domain specialist to be relevant for the analysis, they should also be codified into attributes in the dataset (this is illustrated in the ”Another Attribute” columns in Fig. 4). These domain-specific characteristics may have different granularity levels (associated to an event, to a process instance, or to the process itself).

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA



Instance	Duration of event A	Duration of event B	Duration of event C	Duration of event D	Simultaneous duration of events B and D	Another attribute	...	Class variable
window of analysis 1	2hr	4hr	6hr	8hr	3hr	$50	...	good quality
window of analysis 2	3hr	5hr	7hr	10hr	4hr	$80	...	good quality
window of analysis 3	1hr	3hr	2hr	5hr	1hr	$30	...	poor quality
...	...	...	...	...	...	...	...	...
window of analysis N	1hr	4hr	2hr	7hr	5hr	$60	...	medium quality

Fig. 4 Example of the application of the proposed method, after the ”change point detection” and ”feature engineering” steps

4.3 Feature selection and treatment learning

In the treatment learning step, the combinations of attribute-value pairs (i.e., treatments) that contributed the most to the deviations of the process are discovered. This contribution is evaluated using the *worth* and *lift* metrics, as suggested by Richetti et al. [30].

The *worth* (Equation 1) of a dataset D in terms of distribution of class values is defined as a weighted probability sum of class values in D. The *lift* (Equation 2) of a treatment R_X is the ratio of the *worth* of the treated subset to the worth of the baseline dataset D. A value of $lift(R_X) > 1$ denotes an improvement of the class distribution.

$$worth(D) = \sum_{i=1}^{Classes} (Score(C_i) \times P(C_i)) \quad (1)$$

$$lift(R_x) = \frac{worth(D \cap R_x)}{worth(D)} \quad (2)$$

Hu [16] proposed the TAR3 algorithm for treatment discovery. This algorithm ranks sets of attribute-value pairs (treatments) according to their ability to modify the distribution of the class variable. TAR3 is capable of working with both categorical and continuous variables, and for continuous classes, discretization is performed with an equal-width interval-binning strategy [10]. It starts by mapping all possible values of an attribute to a set of consecutive positive integers. Its core pruning strategy to select the most relevant treatment based on the *confidence1* measure, which measures the difference of an item's confidence in non-best classes with respect to the best class. With ordered class values, the algorithm associates each value with a score, which can also be weighted according to user preference. The class value with the highest score is the best class value C_{best}, and all other values are non-best C_j, ($j \neq best$).

Equation 3 defines the *confidence1* measure, with the following parameters: *a.r* is the value *r* of an attribute *a*, $|a.r|$ is the number of examples in a dataset in which an attribute *a* takes value *r*, $|a.r,C_{best}|$ is the number of examples in which attribute *a* takes value *r* and belongs to class C_{best}, $|a.r,C_j|$ is the number of examples in which attribute *a* takes value *r* and belongs to class C_j, where $j \neq best$ and $S(C_{best})$,$S(C_j)$ are the scores of class C_{best} and C_j, returned by the scoring function.

$$\triangle\, a.r = \frac{\sum_j (S(C_{best}) - S(C_j)(|a.r, C_{best}| - |a.r, C_j|)}{|a.r|} \quad (3)$$

PUC-Rio - Certificação Digital Nº 2020862/CA

TAR3 employs a random sampling strategy assuming that the highest *confidence1* weights tend to have a higher probability of being selected in treatments than those with smaller weights. Where a_i is an attribute-value pair in a dataset and $\triangle i$ is the associated *confidence1* weight of a_i, the strategy samples a_i from a multinomial distribution with discrete weights $\triangle i$. Initially, a_i pairs are then placed in increasing order according to $\triangle i$. Then, the cumulative distribution function (CDF) is calculated for each value of $\triangle i$. A uniform value u in the interval [0,1] is sampled, where the sample is the least a_i such that $u < CDF(i)$. This process is then repeated until a treatment *Rx* of a given size is found. Treatment size is a uniform value sampled from the interval *[1...maxTreatmentSize]*, where *maxTreatmentSize* is the maximum size of interest to a user. The upper bound of *maxTreatmentSize* is the total number of attributes in the dataset. A treatment *Rx* is reported if it is among the top *N* treatments found whose *lift*(*Rx*) > 1, where *N* is the maximum number of treatment the user wants to retrieve.

Since treatments with larger sizes tend to achieve higher lifts than those of smaller sizes (a property known as the universal-existential upward closure) [39], to solve this problem, TAR3 introduces a penalization of *lift(RX)* measurements based on the best class support *sup(Rx)* parameter. *sup(Rx)* is the defined as the ratio of examples in the dataset belonging to the best class in the treated set to those in the complete dataset (Equation 4).

$$sup(R_x) = \frac{|C_{best}, R_x|}{C_{best}} \tag{4}$$

PUC-Rio - Certificação Digital Nº 2020862/CA

A *minSup* parameter specifies the minimum *sup(Rx)* a treatment *Rx* must achieve, and instead of rejecting treatments below this threshold, TAR3 considers *minSup* as a regularizer that penalizes *lift(Rx)* (Equations 5-6). This penalization aims to ensure that potentially highly predictive treatments can be reported even though their *sup(Rx)* are below *minSup*. The stopping criterion of the algorithm involves the following strategy: to generate a set of N treatments, TAR3 performs multiple trials. In each trial, X new treatments are discovered and checked, and only the top N treatments remain in the treatment set. If the current trial does not contribute extra treatments to the top N treatment set, it is called a futile trial. The procedure stops after M futile trials. The complete details of the TAR3 algorithm, its performance evaluation and application in several scenarios can be found in Hu [16].

$$lift(R_x) = \frac{worth\ (D \cap R_x)}{worth\ (D)} \times penalty \tag{5}$$

$$penalty = \begin{cases} 1 & (\sup\ (R_x) \geq minSup) \\ \frac{\sup\ (R_x)}{minSup} & (\sup(R_x) < minSup) \end{cases} \tag{6}$$

Menzies and Hu [26] proposed two types of treatments: controller and monitor, where the former is the treatment that can improve the current baseline situation the most and the latter is the treatment that can degrade the current baseline situation the most. In the loan example, a monitor rule was discovered, since it favored the worst class value after the application of R_1.

According to Hu [16], the effectiveness of TAR3 is reduced when applied to datasets with a large number of variables. Thus, to improve the lift of treatments identified by the algorithm, we propose a feature selection step. The feature selection problem can be formulated as

PUC-Rio - Certificação Digital Nº 2020862/CA

follows: Given an original set *V* of attributes, with cardinality *m* and a subset *x*, $x \subseteq V$, with cardinality *d*, the feature selection criterion for the subset *x* is represented by *J(x)*, so the higher the value of *J(x)* is, the better the subset of *x* attributes will be. Thus, the feature selection problem can be defined by finding a subset $x \subseteq V$, with cardinality $d \leq m$, that solves Equation 7.

$$\begin{gathered} J(x) = \max J(z) \\ s.t.: \\ z \subseteq V \\ d \leq m \end{gathered} \tag{7}$$

An exhaustive search for this problem would require the assessment of all possible $\binom{m}{d}$, for all values of $d \leq m$. According to Yusta [41], the number of possibilities grows exponentially, making an exhaustive search impractical even for moderate values of *m*. For this reason, we propose a metaheuristic to overcome this challenge.

The feature selection criterion *J(x)* used to measure the result of the metaheuristic is the lift of the treatments found after its application. These treatments are parameterized with a maximum size *r*, which can assume the values $r = (1,...,d)$, and will be obtained from a dataset *x*, of cardinality *d*.

The proposed metaheuristic, presented in algorithm 1, looks for treatments with the highest lift *J(x)*, in a dataset *x*, varying the maximum size *r* of treatments until *J(x)* does not increase further with variations of *r* or until only similar redundant treatments of *J(x)* are found. When a stopping criterion is met, the attributes that are not present in the *N* treatments found will be removed from the dataset *x*, according to the value of *r* having the highest *J(x)*. These steps will be repeated until an iteration does not find treatments with a value of *J(x)* higher than the one

PUC-Rio - Certificação Digital Nº 2020862/CA

found in the previous iteration or until the treatments found encompass all attributes that are present in the dataset *x*.

Algorithm 1: Metaheuristic for feature selection

MHFeatureSelection (dataset x, int d): int best_lift, int best_r

1. $J(x)_1$ = Perform treatment learning with dataset *x* and treatment size 1

2. Perform treatment learning with dataset *x* and treatment size *r = (2,...,d)* until $J(x)_r \leq J(x)_{r-1}$ and the size of the largest treatment found with *r* is smaller than *r-1*

3. Record best_lift = $J(x)$ and the treatment size associated with the best lift best_r = r

4. Delete from dataset *x* the attributes that are not present in the treatments discovered with best_r;

5. Rerun until reaching an iteration such that $J(x)_r$ < best_lift for all *r = (1,...,d)* or until reaching a best_r in which the treatments encompass all attributes present in dataset *x*.

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

5 Experimental Results

This section describes the experiments carried out to evaluate the feasibility of the method proposed in Section 4 and the effectiveness gains achieved by the feature selection step presented in Section 4.3. The best treatment obtained by the experiments in the controller view and monitor view are presented, showing their effects on the original distribution of classes (*baseline*) and comparing the results obtained with and without the application of the proposed metaheuristic for feature selection.

5.1 Experimental design

The experiment was conducted using the scenario of the COVID-19 pandemic, with the aim of understanding the most relevant factors that affected the rate of COVID-19 cases in a country, adjusted by its population. We integrated real data from several sources: (i) data about the non-pharmaceutical measures applied by more than 100 countries from all continents, collected from the Oxford COVID-19 government response tracker initiative [13]; (ii) mobility data gathered from mobile devices with location tracking activated [2]; (iii) socioeconomic and demographic indicators for each country. The period of analysis was from the start of the pandemic until May 2021.

Monitoring and combating the COVID-19 pandemic can be understood as a complex process, carried out in different ways by different countries, in which the events can be represented by the non-

[2] https://www.google.com/covid19/mobility/

pharmaceutical and social measures applied to decrease the acceleration of confirmed cases. This process is composed of non-atomic events, since the non-pharmaceutical measures applied are not instantaneous, that is, the applied measures have durations different than zero. Mobility, demographic and socioeconomic indicators also characterize the instances as contextual attributes, and even though they are not as dynamic as the contention measures that were initiated, deactivated and reactivated several times by the public policymakers of each country, they also should be codified as attributes of each instance of the process, since they represent characteristics of the process assumed in each execution.

All the events considered for this process are represented in Fig. 5 and the attributes of the process instances are presented in Fig. 6.

Moreover, it is noteworthy that monitoring and combating the COVID-19 pandemic compose an online process, since these measures were continuously carried out since the outset, and still did not end in any of the executions (countries).

Thus, the chosen scenario is a real online process with non-atomic events, an ideal scenario to validate the application of the proposed method. This work presents insights that can help to better understand the fight against the pandemic, using process mining tools, in particular proposing an innovative metaheuristic approach for deviance mining through treatment learning.

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

Nonpharmaceutical Interventions (NPIs)	Nonpharmaceutical Interventions (NPIs)
No School closing	**C1_0**
Recommend School closing	**C1_1**
Partial School closing	**C1_2**
Require School closing	**C1_3**
No Workplace closing	C2_0
Recommend Workplace closing	C2_1
Partial Workplace closing	C2_2
Require Workplace closing	C2_3
No Cancellation of public events	**C3_0**
Recommend Cancellation of public events	**C3_1**
Require Cancellation of public events	**C3_2**
No Restrictions on gatherings	C4_0
Restrictions on gatherings (> 1000 people)	C4_1
Restrictions on gatherings (100 > people < 1001)	C4_2
Restrictions on gatherings (10 > people < 101)	C4_3
Restrictions on gatherings (people < 11)	C4_4
No Closing of public transport	**C5_0**
Recommend Closing or significant reduction of public transport	**C5_1**
Required Closing of public transport	**C5_2**
No Stay at home requirements	C6_0
Stay at home recommendation	C6_1
Just leaving house for essential trips	C6_2
Require not leaving house with minimal exceptions	C6_3
No restrictions on internal movement	**C7_0**
Recommend not to travel between regions/cities	**C7_1**
Internal movement restrictions in place	**C7_2**
No restrictions to international travels	C8_0
Screening arrivals	C8_1
Quarantine arrivals from some or all regions	C8_2
Ban arrivals from some regions	C8_3
Total border closure	C8_4
No Income support	**E1_0**
Government is replacing less than 50% of lost salary	**E1_1**
Government is replacing 50% or more of lost salary	**E1_2**
No debt/contract relief for households	E2_0
Narrow relief (specific to one kind of contract)	E2_1
Broad debt/contract relief	E2_2
No Covid-19 public information campaign	**H1_0**
Public officials urging caution about Covid-19	**H1_1**
Coordinated public information campaign	**H1_2**
No testing policy	H2_0
Only testing those who both have symptoms and meet specific criteria	H2_1
Testing of anyone showing Covid-19 symptoms	H2_2
Open public testing	H2_3
No contact tracing	**H3_0**
Limited contact tracing	**H3_1**
Contact tracing for all identified cases	**H3_2**
No policy for facial covering	H6_0
Recommended facial covering	H6_1
Required facial covering in some shared public spaces	H6_2
Required facial covering in all shared public spaces	H6_3
Required facial covering in all public spaces	H6_4
No Vaccination Policy	**H7_0**
Vaccination available for ONE of following: Key workers/ clinically vulnerable groups (non elderly)/ elderly groups	**H7_1**
Vaccination available for TWO of following: Key workers/ clinically vulnerable groups (non elderly)/ elderly groups	**H7_2**
Vaccination available for ALL of following: Key workers/ clinically vulnerable groups (non elderly)/ elderly groups	**H7_3**
Vaccination available for all three plus partial additional availability	**H7_4**
Universal availability	**H7_5**
No policies for protecting elderly people	H8_0
Recommended isolation	H8_1
Narrow restrictions for isolation	H8_2
Extensive restrictions for isolation	H8_3

Fig. 5 Dataset attributes and corresponding codes referring to the Non-pharmaceutical Interventions carried out by several countries in response to the COVID-19 outbreak

PUC-Rio - Certificação Digital Nº 2020862/CA

Data Type	Data
Demographic Data	Percentage of population living in urban areas
Demographic Data	# of cities with population larger than 17M
Demographic Data	Average population density of cities with population larger than 17M (hab/m²)
Demographic Data	# of cities with population larger than 13M and less than 17M
Demographic Data	Average population density of cities with population larger than 13M and less than 17M (hab/m²)
Demographic Data	# of cities with population larger than 9M and less than 13M
Demographic Data	Average population density of cities with population larger than 9M and less than 13M (hab/m²)
Demographic Data	# of cities with population larger than 1M and less than 9M
Demographic Data	Average population density of cities with population larger than 1M and less than 9M (hab/m²)
Demographic Data	# of cities with population larger than 500K and less than 1M
Demographic Data	Average population density of cities with population larger than 500K and less than 1M (hab/m²)
Socioeconomic Data	GDP per Capita (US$)
Mobility Data	Retail and recreation percent change from baseline
Mobility Data	Grocery and pharmacy percent change from baseline
Mobility Data	Parks percent change from baseline
Mobility Data	Transit stations percent change from baseline
Mobility Data	Workplaces percent change from baseline
Mobility Data	Residential percent change from baseline
Demographic Data	Median age
Demographic Data	Life expectancy
Class Variable	Acceleration of the contaminated rate

Fig. 6 Dataset attributes referring to other relevant perspectives for the process of combating COVID-19

The second derivative of the number of confirmed COVID-19 cases in each country (acceleration of the number of infected individuals) was selected to be the class variable of this process, that is, the treatments discovered are evaluated according to their effects on the acceleration of the number of infected individuals. The class variable was discretized to four possible values: ”high acceleration”, ”medium-high acceleration”, ”medium-low acceleration” and ”low acceleration”. The values were defined by applying a quartile analysis to each country, considering all the values of the acceleration of the number of infected people of that specific country during the analysis period. Since different quartiles were defined for each country, the discretization of the class variable did not suffer any possible bias.

The dataset for the treatment learning step of the proposed method was then codified, taking as input all collected and integrated data about the online process. To define the process instances of the new dataset, change point detection was applied. This step sought structural changes in the time series of the class variable of each country following the definitions of Zeileis et al. [44]. With the identification of change points, each analysis window was defined as the period between a first change point and its subsequent one, and included data from non-pharmaceutical measures, demographic and socioeconomic characteristics, and mobility in that period. To identify the change points in the time series of the class variable, the R system [3] **strucchange** package [43] was used.

Each window of analysis was associated with the value of the class variable on the last day present in that analysis window. That is, as an illustrative example, where an analysis window for Brazil contained data from March 15, 2020 to August 25, 2020, and on August 25th the class variable in Brazil was characterized as ”high acceleration”, this instance’s class variable was set to ”high acceleration”.

After defining the windows, the instances of the new dataset were defined and characterized in terms of their features. The feature engineering step aims to choose attributes that enable a good representation of process data in the new dataset; since it is a process with non-atomic events, the temporal dimension of the events needs to be represented. Thus, two temporary datasets were created, one containing data on the duration of non-pharmaceutical measures in each analysis window, called ”Single Non-pharmaceutical Interventions” (SNPI), and a second dataset containing data on the simultaneous duration of all possible pairs of non-pharmaceutical measures in each analysis window, called ”Non-pharmaceutical Intervention Sets” (NPIS). The duration time and simultaneous duration time are represented as percentages of the total time of the analysis window in which they are present. In both datasets, the attributes related to demographic, mobility and socioeconomic indicators were also represented, with a fixed value set to the value that

[3] https://www.r-project.org/

PUC-Rio - Certificação Digital Nº 2020862/CA

each variable assumed in the period and in the country of each analysis window.

The feature engineering step generated a large number of attributes in each dataset: the Single Non-pharmaceutical Interventions dataset contained a total of 84 attributes and the Non-pharmaceutical Intervention Sets dataset contained 1817 attributes. To improve the effectiveness of the treatment learning step, the feature selection step was set to reduce the dimensionality, by removing attributes that were not significant for the process result. After applying the feature selection metaheuristic, the Single Non-pharmaceutical Interventions dataset was reduced to 13 attributes in the Monitor experiment and 18 attributes in the Controller experiment, and the Non-pharmaceutical Intervention Sets dataset was reduced to 20 attributes in the Monitor experiment and 15 attributes in the Monitor experiment Controller experiment. The effectiveness gains generated by this step of the method are presented in Section 5.7.

Two experiments were carried out for each dataset in the treatment learning step, one looking for the sets of attributes that worsened the distribution of the class variable (Monitor) and a second looking for the sets of attributes that improved the distribution of the class variable (Controller). Each experiment searched for the 30 treatments that caused the most significant changes. The results of this step are presented in Sections 5.2, 5.3, 5.4 and 5.5.

Additionally, we assessed the benefit of considering non-atomic events (representing the non-pharmaceutical interventions) in the dataset codification, in comparison with a dataset composed exclusively of atomic events. The results of this assessment are in Section 5.6.

The results of the method were evaluated in two rounds. First, we conducted an interview with an epidemiologist and researcher on COVID-19 to assess the quality and applicability of the treatments found,

PUC-Rio - Certificação Digital Nº 2020862/CA

as described in Section 5.8. Then, the interpretability and effectiveness of the discovered rules were assessed through an experiment, as described in Section 5.9. The data from the experiments can be found at: https://github.com/LJazbik/Mestrado.git.

5.2 Monitor SNPI treatments

Fig. 7 shows the 30 treatments that were found in the Monitor SNPI experiment, with their corresponding worth distributions.

Fig. 7 Worth of each treatment discovered in the 'Monitor SNPI' experiment

The best monitor treatment considering Single Non-pharmaceutical Interventions (T1 Monitor SNPI) was the following, with a worth of 1.819587:

Number of cities with population between 13M and 9M inhab.= [0..1) AND [C3_2=[0.873315..1] AND C2_1=[0.000000..0.003571) AND [C2_0=[0..0.011287) AND [C4_2=[0..0.003040) AND [C3_0=[0..0.004717)]

PUC-Rio - Certificação Digital Nº 2020862/CA

This treatment indicates that the analysis windows of countries with no cities having populations between 13M and 9M and with mandatory cancellation of public (from 87.3315% to 100%), closing of schools or operation of schools with special prevention measures (from 0% to 0.3571%), without closure of working environments (from 0% to 1.1287%), with restriction of agglomerations between 101 and 1000 people (from 0% to 0.304%) and no cancellation of public events (from 0% to 0.4717%) impacted the class variables with the distribution shown in Table 2.

Table 2 Distribution change among class values for the T1 Monitor SNPI treatment

Class values	*Baseline*	*After treatment*
high acceleration	22%	64%
medium-high acceleration	4%	7%
medium-low acceleration	52%	11%
low acceleration	22%	18%

The application of the best Monitor treatment considering Single Non-pharmaceutical Interventions resulted in an increase in the acceleration of the COVID-19 infection rate. As shown by Table 2, there was an increase from 22% to 64% in the number of analysis windows with high acceleration and an increase from 4% to 7% in the number of analysis windows with - acceleration. At the same time, there were decreases from 52% to 11% and 22% to 18% in the number of analysis windows with medium-low and llo acceleration, respectively.

PUC-Rio - Certificação Digital Nº 2020862/CA

5.3 Controller SNPI treatments

Fig. 8 shows the 30 treatments that were found in the Controller SNPI experiment, with their corresponding worth distributions.

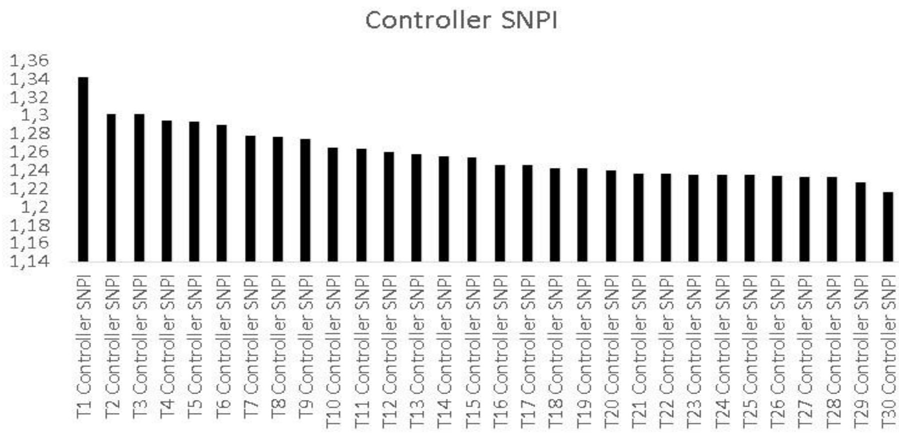


Fig. 8 Worth of each treatment discovered in the 'Controller SNPI' experiment

The best Controller treatment considering Single Non-pharmaceutical Interventions, T1 Controller SNPI, was the following, with a worth of 1.341618:

[H2_0=[0..0.002342) AND H8_3=[0..0.011461) AND C6_3=[0..0.005698) AND C2_3=[0..0.009217)];

This treatment indicates that the analysis windows with no testing policy (from 0% to 0.2342%), severe restrictions for isolation of elderly people (from 0% to 0.11461%), requirement of facial coverings in all shared/public spaces (from 0% to 0.5698%) and required closing of all non-essential workplaces (from 0% to 0.9217%) impacted the class variables with the distribution shown in Table 3.

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

Table 3 Distribution change among class values for the T1 Controller SNPI treatment

Class values	*Baseline*	*After treatment*
high acceleration	22%	23%
medium-high acceleration	4%	2%
medium-low acceleration	52%	17%
low acceleration	22%	58%

The application of the best Controller treatment considering Single Non-pharmaceutical Interventions resulted, in most cases, in a reduction in the acceleration of the COVID-19 infection rate, as shown in Table 3. Although there was a slight increase from 22% to 23% in the number of analysis windows with high acceleration, there was a reduction from 4% to 2% in the number of analysis windows with medium-high acceleration and a reduction from 52% to 17% in the number of windows with medium-low acceleration, along with an increase from 22% to 58% in the number of windows with low acceleration.

5.4 Monitor NPIS treatments

Fig. 9 shows the 30 treatments that were found in the Monitor NPIS experiment, with their corresponding worth distributions.

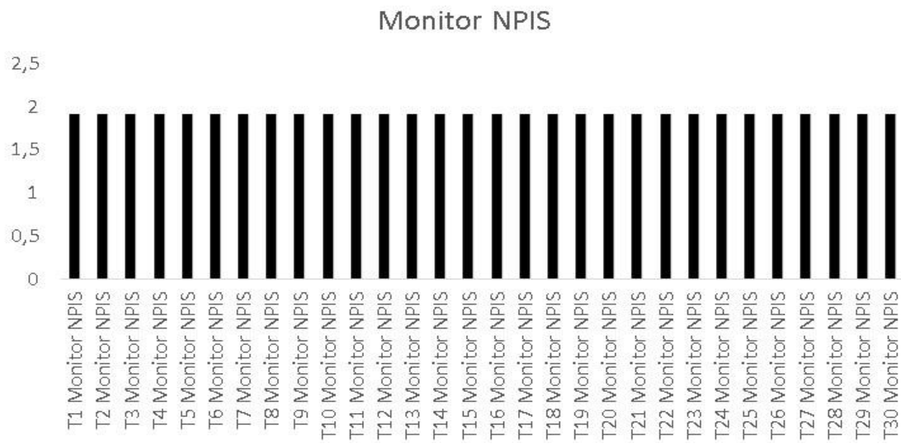


Fig. 9 Worth of each treatment discovered in the Monitor NPIS experiment

PUC-Rio - Certificação Digital Nº 2020862/CA

The best monitor treatment considering Non-pharmaceutical Intervention Sets, T1 Monitor NPIS, was the following, with a worth of 1.914864:

[E2 _0 AND H8_0=[0..0.009756)] AND [C1_0 AND H_ 3=[0..0.015707)] AND [H2_0 AND H7_ 0=[0..0.002342)] AND [C1_0 AND H1_0=[0..0.002252)] AND [H1_0 AND H3_0=[0..0.002252)] AND [C5_0 AND E2_0=[0..0.002451)] AND [C4_0 AND H8_0=[0..0.009756)] AND [C4_0 AND H1_0=[0..0.002252)] AND [C2_0 AND H6_0=[0..0.012500)] AND [C5_0 AND H2_0=[0..0.002342)] AND [C8_4 AND H2_3=[0..0.007246)] AND [C3_2 AND C4 _4=[0.647059..1]] AND [C4 _0 AND C7_0=[0..0.049550)] AND [C4_0 AND C5_0=[0..0.077273)] AND [C4_0 AND E2_0=[0..0.011287)];

This treatment denotes that the analysis windows with no household debt relief and severe restrictions for isolation of elderly people (from 0% to 0.9756%), no school closing and with vaccination available to key workers, clinically vulnerable and elderly people (from 0% to 1.5705%), no testing policy and no vaccination policy (from 0% to 0.2342%), no school closing and no public information campaigns (from 0% to 0.2252%), no public information campaigns and no contact tracing after positive diagnosis (from 0% to 0.2252%), no closing of public transport and no or narrow household debt relief (from 0% to 0.2451%), no closing of public transport and no household debt relief (from 0% to 0.2451%), no restrictions on gatherings and no restrictions for isolation of elderly people (from 0% to 0.9756%), no restrictions on gatherings and no schools closing (from 0% to 0.2252%), no workplace closing and no facial covering policy (from 0% to 1.25%), no closing of public transport and no testing policy (from 0% to 0.2342%), total border closure and open public COVID-19 testing (from 0% to 0.7246%), with cancelation of public events and restrictions on gatherings of 10 people or fewer (from 64.7059% to 100%), with required cancelling public events and restrictions on gatherings of 10 people or fewer (from 64.7059% to 100%), no restrictions on gatherings and no restrictions on internal movement between cities/regions (from 0% to 4.955%), no restrictions on internal movement between cities/regions and no closing of public transport (from 0% to 7.7273%), no restrictions on internal movement between cities/regions and no or narrow household debt relief (from 0% to 1.1287%) impacted the class variables with the distribution shown in Table 4.

Table 4 Distribution change among class values for the T1 Monitor NPIS treatment

Class values	*Baseline*	*After treatment*
high acceleration	22%	70%
medium-high acceleration	4%	5%
medium-low acceleration	52%	3%
low acceleration	22%	23%

PUC-Rio - Certificação Digital Nº 2020862/CA

The application of the best Monitor treatment considering Non-pharmaceutical Intervention Sets resulted in an increase in the acceleration of the COVID-19 infection rate in most cases, as shown by Table 4. There was an increase from 22% to 70% in the number of analysis windows with high acceleration and an increase from 4% to 5% in the number of analysis windows with medium-high acceleration. At the same time, there was a decrease from 52% to only 3% in the number of windows with medium-low acceleration and only a slight increase, from 22% to 23%, in the number of analysis windows with low acceleration.

PUC-Rio - Certificação Digital Nº 2020862/CA

5.5 Controller NPIS treatments

Fig. 10 shows the 30 treatments that were found in the Controller NPIS experiment, with their corresponding worth distributions.

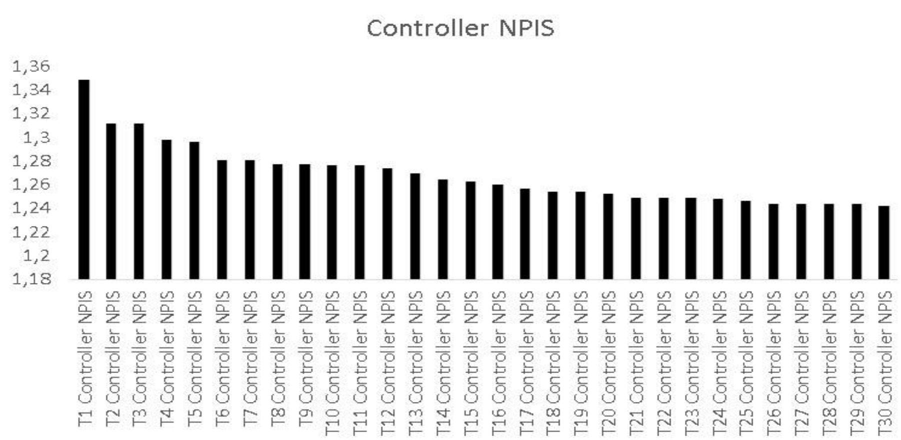


Fig. 10 Worth of each treatment discovered in the Controller NPIS experiment

The best Controller treatment considering Non-pharmaceutical Intervention Sets, T1 Controller NPIS, was the following, with a worth of 1.347999:

[C4_4 AND C8_3=[0..0.002252)] AND [C2_2 AND H3_1=[0..0.006757]] AND [C4_4 AND C5_1=[0..0.003401)];

This treatment indicated that the analysis windows with restrictions on gatherings of 10 people or fewer and travel controls for some countries (from 0% to 0.2252%), with required workplace closing for some sectors or categories and limited contact tracing after positive diagnosis (from 0% to 0.6757%), and with restrictions on gatherings of 10 people or fewer and closing or significantly reduced volume of transport available (from 0% to 0.3401%) impacted the class variables with the distribution shown in Table 5.

Table 5 Distribution change among class values for the T1 Controller NPIS treatment

Class values	*Baseline*	*After treatment*
high acceleration	22%	12%
medium-high acceleration	4%	2%
medium-low acceleration	52%	36%
low acceleration	22%	50%

The application of the best Controller treatment considering Non-pharmaceutical Intervention Sets resulted in a reduction in the acceleration of the COVID-19 infection rate, as shown by Table 5. There was a reduction from 22% to 12% in the number of analysis windows with high acceleration, a reduction from 4% to 2% in the number of analysis windows with medium high acceleration, a reduction of 52% to 36% in the number of analysis windows with medium low acceleration and a big increase from 22% to 50% in the number of analysis windows with low acceleration.

PUC-Rio - Certificação Digital Nº 2020862/CA

5.6 Assessing the benefit of codifying non-atomic events

Additionally, we assessed the benefit of considering non-atomic events (representing the non-pharmaceutical interventions) in the dataset's codification, when compared to a dataset composed exclusively of atomic events. The results of this assessment are in Section 5.6.

We also conducted an experiment to specifically assess the benefit of using non-atomic events to represent activities in the dataset.

The purpose of this section is to verify if coding activities as non-atomic has any advantage over coding them as atomic. For this, we compared the results of the treatment learning application on the SNPI dataset and on an adaptation of the same, in which the non-pharmaceutical measures to combat the pandemic were coded as atomic. While the SNPI dataset presents the non-pharmaceutical measures to combat the pandemic in relation to their duration, the adapted dataset presents such measures only in relation to their existence or not in the windows of analysis, so that the measures are represented by a binary variable where ”0” represents the non-existence and ”1” represents the existence.

The treatment learning stage discovered treatments with greater worth for the Monitor and Controller views, when performed on the dataset with non-atomic measurements (SNPI dataset), as shown by Figures 11 and 12.

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

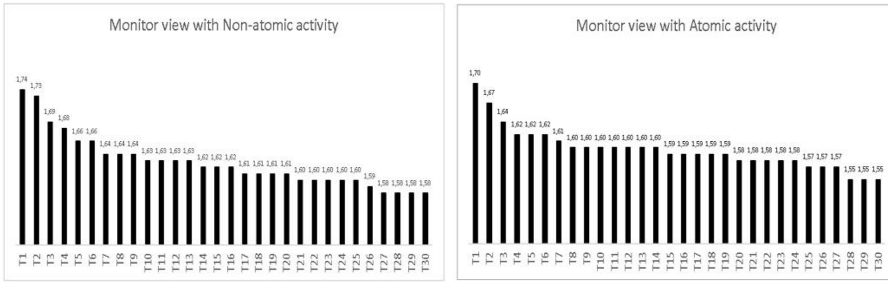


Fig. 11 Worth of each treatment discovered in the Monitor view with non-atomic and atomic data

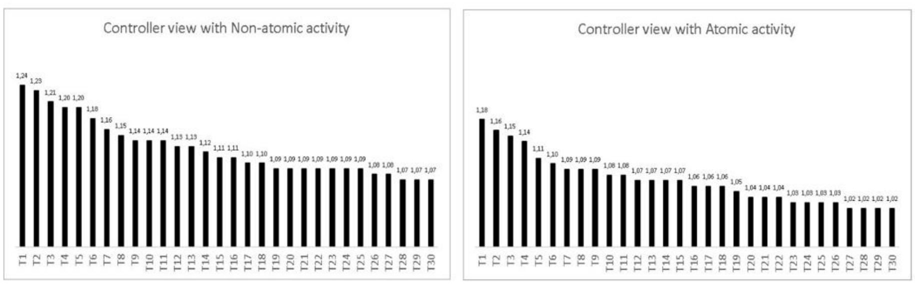


Fig. 12 Worth of each treatment discovered in the Controller view with non-atomic and atomic data

5.7 Evaluating the proposed metaheuristic

In this section, the resulting treatments obtained by the method using the metaheuristic are compared to the execution of the method without the metaheuristic. The comparison was performed both for the Controller and Monitor views.

As can be seen in Figures 13, 14, 15 and 16, in all cases the application of the proposed metaheuristic led to the discovery of treatments with higher worth. Thus, the use of the metaheuristic led to an increase in effectiveness of the treatment learning step, regardless of the number of best treatments considered.

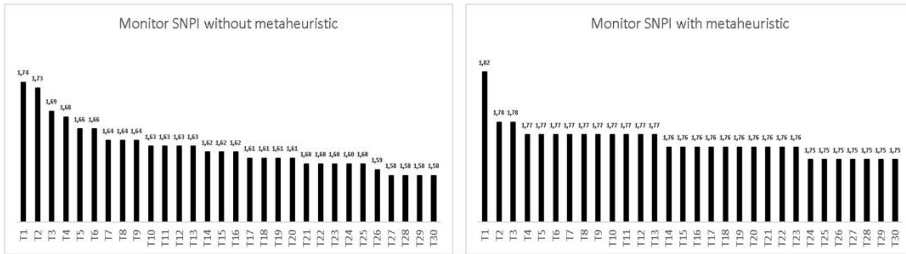


Fig. 13 Worth of each treatment discovered in the Monitor SNPI experiment with (right) and without (left) the application of the metaheuristic

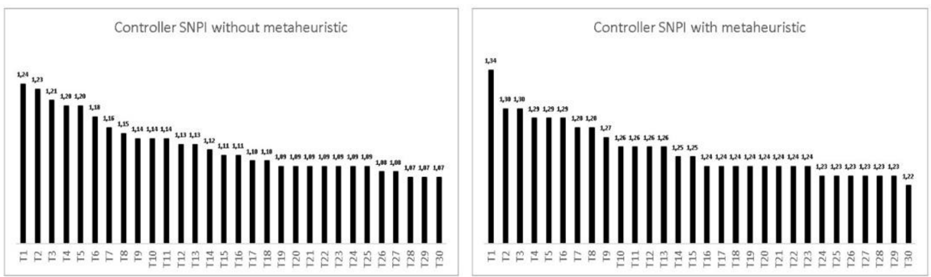


Fig. 14 Worth of each treatment discovered in the Controller SNPI experiment with (right) and without (left) the application of the metaheuristic

Fig. 15 Worth of each treatment discovered in the Monitor NPIS experiment with (right) and without (left) the application of the metaheuristic

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

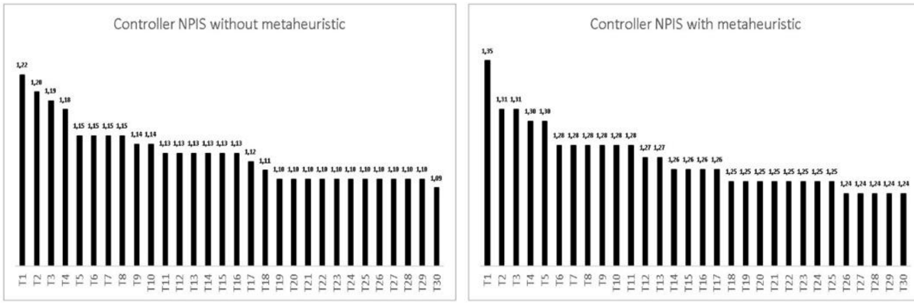


Fig. 16 Worth of each treatment discovered in the Controller NPIS experiment with (right) and without (left) the application of the metaheuristic

5.8 Qualitative analysis by the domain specialist

The treatments found by the proposed method were presented to an epidemiologist and researcher in the field of COVID-19 so that he could analyze the quality and applicability of the treatments found in policies to combat COVID-19.

This presentation detailed the dynamics of the treatment learning stage, the variables considered as process events, the data organization structure, the experiments carried out with a focus on the duration time and the simultaneity of occurrence of the events, the period of the pandemic that the data represented and the results obtained.

The results showed that the cancellation of events and the restriction of agglomerations, which were the first measures to be taken by governments, were not enough to contain the pandemic.

After the presentation, a semi-structured interview was conducted to collect feedback on the results found and the hypotheses formulated from them. This interview included the following 4 (four) questions:

1. “What is your opinion on this hypothesis?”
2. “What is your opinion regarding the effectiveness of the proposed method?”

3. “Is the duration of non-pharmaceutical measures relevant to be analyzed?”
4. “Is the simultaneous occurrence of non-pharmaceutical measures relevant to be analyzed?”

The expert answers confirmed our research hypothesis on the relevance of taking the duration and simultaneity of occurrence of the events into account. Regarding the effectiveness of the proposed method, he agreed that the method was effective.

The expert also commented that the treatments discovered were too complex and difficult to be interpreted by someone unfamiliar with the logical language in which they are described. Therefore, seeking to increase the interpretability of the results obtained by the proposed method, a new experimental scenario was configured as described in Section 5.9.

PUC-Rio - Certificação Digital Nº 2020862/CA

5.9 Assessing the treatments interpretability

In this section, we extended our proposed method and evaluate the gains in interpretability of the treatments found.

In this extended version of our method, we consider that treatments are more readable (and therefore more interpretable) when their formal expression does not contain negative terms, since: “negative statements are psychologically more complex and harder to process” [53,54]. Thus, a new hypothesis was defined to direct the treatments found by the method: that the negative variables (those that represent the non-occurrence of an event in a process instance, for example, “No school Closing”) should be avoided, even if this would lead to a reduction in the effectiveness of the results.

Therefore, the feature engineering step of the proposed method was adjusted to disregard features that represent the non-occurrence of a non-pharmaceutical measure of combat the pandemic (that is, only attributes that represent the occurrence of some event were created).

The best NPIS treatments found in this experiment scenario, respectively in the monitor view and in the controller view, were:

NPIS_Monitor:*[[C3_2 AND H1_2 = [0.859649..1.0]] AND [C8_1 AND H3_2 = [0..0.007772]] AND [C8_4 AND H8_3 = [0..0.005731]];*

NPIS_Controller:*[[C2_2 AND C3_2 = [0..0.187654]] AND [C4_4 AND H8_2 = [0..0.002577]];*

The interpretation of the NPIS_Monitor treatment is that:

- cancellation of public events and conducting an awareness campaign in a coordinated manner occurred simultaneously during 85% to 100% of the time,
- elderly isolation recommendation and contact tracing for all positive cases occurred simultaneously during only 0% to 0.08% of the time, and
- total border closures and strict restrictions on the isolation of the elderly occurred simultaneously during only 0% to 0.06% of the time.

In turn, the interpretation of the NPIS_Controller treatment is that:

- partial closure of workplaces and cancellation of public events occurred simultaneously during only 0% to 0.19% of the time and

PUC-Rio - Certificação Digital Nº 2020862/CA

- restrictions on gatherings of more than 11 people and isolation of the elderly occurred simultaneously during only 0% to 0.02% of the time.

The NPIS_Monitor treatment obtained a worth of 1.599368 and the NPIS_Controller treatment obtained a worth of 1.228281. When we compared these worth values with the ones obtained by the previous NPIS experiment, there was a worth loss of 0.13 in the monitor view, while worth values were similar in the controller view.

When analyzing the effects of excluding the attributes of non-occurrence in the SNPI experiment, the results in the monitor and controller views were, respectively :

SNPI_Monitor: *[[C2_1=[0..0.003571]] AND [C3_2=[0.952703..1]];*

SNPI_Controller: *[[C3_2=[0..0.445946]];*

The interpretation of the NPIS_Monitor treatment is that:

- workplace closure recommendation almost never occurred (0% to 0.3% of the time) and
- cancellation of public events almost always occurred (95% to 100% of the time).

And the interpretation of the NPIS_Controller treatment is that:

- cancellation of events with audiences occurred 0% to 44.6% of the time.

The SNPI_Monitor treatment obtained a Worth of 1.727318 and the SNPI_Controller treatment obtained a Worth of 1.229962. When comparing these values with the results of the previous NPIS experiment,

PUC-Rio - Certificação Digital Nº 2020862/CA

the worth slightly decreased in both views (0.02 in the monitor view and 0.01 in the controller view).

Such results demonstrate that the exclusion of negative attributes leads to results with lower worths, but that are easier to interpret. Thus, it is up to the user of the proposed method (i.e., the process manager) to trade-off between the worth and the interpretability of the treatments to be discovered, analyzing whether in his particular domain it is more relevant to obtain more effective or more interpretable treatments.

PUC-Rio - Certificação Digital Nº 2020862/CA

PUC-Rio - Certificação Digital Nº 2020862/CA

6 Conclusion

The purpose of deviance mining is to identify deviations in the execution of a process in relation to expected behavior and characterize these deviations as positive or negative in relation to their effects on the outcome of the process. It is sub-area of process mining, capable of generating several insights for process analysts and executors. However, existing works on deviance mining do not address the challenges posed by online scenarios and by the presence of non-atomic events. This work proposed a method for online deviance mining of processes with non-atomic events, and an innovative metaheuristic approach for its feature selection step. The proposed method is a domain-driven approach, since all the dataset attributes are built taken into account a specific domain.

The experimental results proved the feasibility of the method in a real-life scenario of the domain of the COVID-19 pandemic, and increased the effectiveness of treatment learning provided by the proposed metaheuristic, enabling discovery of more significant treatments in relation to changes in the outcome of the processes.

In particular, this work exploited real-life data and discovered insights on a very relevant and current domain that has been challenging researchers from around the world from many research areas regarding monitoring and combating the COVID-19 pandemic. To the best of our knowledge, no other studies approached the global fight against the COVID-19 pandemic as a process and applied mining techniques to extract knowledge from this process. This scenario was chosen because it is a real and complex online process with non-atomic events, thus making it ideal for demonstrating the feasibility and flexibility of the proposed method. The application of the proposed method to this scenario proved the method’s feasibility and effectiveness, the performance gains caused by the proposed metaheuristic, and the possibility of adapting the proposed method to discover process treatments that are easier to interpret.

The social impact of the proposed research lies in the benefits of bringing process mining applications closer to real life by considering the temporal dimension of process events and obtaining more explainable results.

Regarding the limitations of this research, we aim to apply the method to other domains for further evaluation to address its generality.

In future work, we intend to examine different temporal relations, assess the impact of several ranges for the simultaneous duration of pairs of events, and address other relationships between pairs of events (for example, the ones provided by declare language) . In addition, we intend to extend the feature selection step with other metaheuristic approaches.

PUC-Rio - Certificação Digital Nº 2020862/CA

References

[1] M. L. Bernardi, M. Cimitile, C. di Francescomarino, and F. M. Maggi. Using discriminative rule mining to discover declarative process models with non-atomic activities. In *Rules on the Web. From Theory to Applications*, pages 281–295. Springer International Publishing, 2014. DOI: 10.1007/ 978-3-319-09870-8 21. URL https://doi.org/10.1007/978-3-319-09870-8 21.

[2] M. L. Bernardi, M. Cimitile, C. di Francescomarino, and F. M. Maggi. Do activity lifecycles affect the validity of a business rule in a business process? Information Systems, 62: 42–59, Dec. 2016. DOI: 10.1016/j.is.2016.06. 002. URL https://doi.org/10.1016/j.is.2016.06.002.

[3] R. J. C. Bose and W. M. van der Aalst.
Discovering signature patterns from event logs. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, Apr. 2013. DOI: 10.1109/cidm.2013. 6597225. URL https://doi.org/10.1109/ cidm.2013.6597225.

[4] Z. D. Bozorgi, I. Teinemaa, M. Dumas, M. La Rosa, and A. Polyvyanyy. Prescriptive process monitoring for cost-aware cycle time reduction. In *2021 3rd International Conference on process mining (ICPM)*, pages 96–103. IEEE, 2021.

[5] A. Burattin, M. Cimitile, F. M. Maggi, and A. Sperduti. Online discovery of declarative process models from event streams. *IEEE Transactions on Services Computing*, 8(6): 833–846, Nov. 2015. DOI: 10.1109/tsc.2015. 2459703. URL https://doi.org/10.1109/tsc.2015.2459703.

[6] J. Carmona and R. Gavald`a. Online techniques for dealing with concept drift in process mining. In *Advances in Intelligent Data Analysis XI*, pages 90–102. Springer
Berlin Heidelberg, 2012. DOI: 10.1007/ 978-3-642-34156-4 10. URL https://doi.org/ 10.1007/978-3-642-34156-4 10.

[7] P. Ceravolo, G. M. Tavares, S. B. Junior, and E. Damiani. Evaluation goals for online process mining: a concept drift perspective. *IEEE Transactions on Services Computing*, pages 1–1, 2020. DOI: 10.1109/tsc.2020. 3004532. URL https://doi.org/10.1109/tsc.2020.3004532.

PUC-Rio - Certificação Digital Nº 2020862/CA

[8] H. J. Cheng, C. Ou-Yang, and Y. C. Juan. A hybrid evolutionary-based process mining technology to discover parallelism structures. In *2012 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, Dec. 2012. DOI: 10.1109/ieem.2012.6838011. URL https://doi.org/10.1109/ieem.2012.6838011.

[9] A. del R´ıo-Ortega, M. Resinas, C. Cabanillas, and A. Ruiz-Cort´es. On the definition and design-time analysis of process performance indicators. *Information Systems*, 38(4):470– 490, 2013.

[10] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995*, pages 194–202. Elsevier, 1995. DOI: 10.1016/b978-1-55860-377-6. 50032-3. URL https://doi.org/10.1016/ b978-1-55860-377-6.50032-3.

[11] S. Dunzer, M. Stierle, M. Matzner, and S. Baier. Conformance checking. In *Proceedings of the 11th International Conference on Subject-Oriented Business Process Management - S-BPM ONE '19*. ACM Press, 2019. DOI: 10.1145/3329007.3329014. URL https://doi.org/10.1145/3329007.3329014.

[12] F. Folino, G. Greco, A. Guzzo, and L. Pontieri. Mining usage scenarios in business processes: Outlier-aware discovery and runtime prediction. *Data & Knowledge Engineering*, 70(12):1005–1029, Dec. 2011. DOI: 10.1016/j.datak.2011.07.002. URL https://doi.org/10.1016/j.datak.2011.07.002.

[13] T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, and H. Tatlow. A global panel database of pandemic policies (oxford COVID-19 government response tracker). *Nature Human Behaviour*, 5(4):529–538, Mar. 2021. DOI: 10.1038/s41562-021-01079-8. URL https:// doi.org/10.1038/s41562-021-01079-8.

[14] M. Hassani. Concept drift detection of event streams using an adaptive window. In *ECMS 2019 Proceedings edited by Mauro Iacono, Francesco Palmieri, Marco Gribaudo, Massimo Ficco*. ECMS, June 2019. DOI: 10.7148/ 2019-0230. URL https://doi.org/10.7148/ 2019-0230.

[15] M. Hassani, S. Siccha, F. Richter, and T. Seidl. Efficient process discovery from event streams using sequential pattern mining. In *2015 IEEE Symposium Series on*

PUC-Rio - Certificação Digital Nº 2020862/CA

Computational Intelligence. IEEE, Dec. 2015. DOI: 10.1109/ssci.2015.195. URL https://doi. org/10.1109/ssci.2015.195.

[16] Y. Hu. Treatment learning : implementation and application. 2003. DOI: 10.14288/1. 0065385. URL https://doi.library.ubc.ca/10.14288/1.0065385.

[17] V. Kumar and S. Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

[18] W. L. J. Lee, A. Burattin, J. Munoz-Gama, and M. Sepu´lveda. Orientation and conformance: A HMM-based approach to online conformance checking. *Information Systems*, 102:101674, Dec. 2021. DOI: 10.1016/j.is.2020. 101674. URL https://doi.org/10.1016/j.is.2020.101674.

[19] G. Li and W. M. van der Aalst. A framework for detecting deviations in complex event logs. *Intelligent Data Analysis*, 21 (4):759–779, Aug. 2017. ISSN 1088467X, 15714128. DOI: 10.3233/IDA-160044. URL https://doi.org/10.3233/IDA-160044.

[20] W. Li, Y. Fan, W. Liu, M. Xin, H. Wang, and Q. Jin. A self-adaptive process mining algorithm based on information entropy to deal with uncertain data. *IEEE Access*, 7:131681–131691, 2019. DOI: 10.1109/access. 2019.2939565. URL https://doi.org/10.1109/ access.2019.2939565.

[21] N. Liu, J. Huang, and L. Cui. A framework for online process concept drift detection from event streams. In *2018 IEEE International Conference on Services Computing (SCC)*. IEEE, July 2018. DOI: 10.1109/scc. 2018.00021. URL https://doi.org/10.1109/scc.2018.00021.

[22] A. Maaradji, M. Dumas, M. L. Rosa, and A. Ostovar. Detecting sudden and gradual drifts in business processes from execution traces. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2140–2154, Oct. 2017. DOI: 10.1109/tkde.2017.2720601. URL https://doi.org/10.1109/tkde.2017.2720601.

[23] A. E. Marquez-Chamorro, M. Resinas, and A. Ruiz-Cortes. Predictive monitoring of business processes: A survey. *IEEE Transactions on Services Computing*, 11(6):962– 977, Nov. 2018. DOI: 10.1109/tsc.2017. 2772256. URL https://doi.org/10.1109/tsc.2017.2772256.

PUC-Rio - Certificação Digital Nº 2020862/CA

[24] J. Martjushev, R. P. J. C. Bose, and W. M. P. van der Aalst. Change point detection and dealing with gradual and multi-order dynamics in process mining. In *Lecture Notes in Business Information Processing*, pages 161–178. Springer International Publishing, 2015. DOI: 10.1007/ 978-3-319-21915-8 11. URL https://doi.org/ 10.1007/978-3-319-21915-8 11.

[25] T. Menzies and Y. Hu. Computing practices data mining for very busy people. *Computer*, 36(11):22–29, Nov. 2003. DOI: 10.1109/mc. 2003.1244531. URL https://doi.org/10.1109/ mc.2003.1244531.

[26] T. Menzies and Y. Hu. Just enough learning (of association rules): the TAR2 “treatment” learner. *Artificial Intelligence Review*, 25(3):211–229, May 2006. DOI: 10.1007/ s10462-007-9055-0. URL https://doi.org/10. 1007/s10462-007-9055-0.

[27] N. Navarin, M. Cambiaso, A. Burattin, F. M. Maggi, L. Oneto, and A. Sperduti. Towards online discovery of data-aware declarative process models from event streams. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2020. DOI: 10. 1109/ijcnn48605.2020.9207500. URL https: //doi.org/10.1109/ijcnn48605.2020.9207500.

[28] H. Nguyen, M. Dumas, M. L. Rosa, F. M. Maggi, and S. Suriadi. Business process deviance mining: Review and evaluation, 2016.

[29] A. Ostovar, A. Maaradji, M. L. Rosa, A. H. M. ter Hofstede, and B. F. V. van Dongen. Detecting drift from event streams of unpredictable business processes. In *Conceptual Modeling*, pages 330–346. Springer International Publishing, 2016. DOI: 10.1007/ 978-3-319-46397-1 26. URL https://doi.org/ 10.1007/978-3-319-46397-1 26.

[30] P. H. P. Richetti, L. S. Jazbik, F. A. Bai˜ao, and M. L. M. Campos. Deviance mining with treatment learning and declare-based encoding of event logs. *Expert Systems with Applications*, 187:115962, Jan. 2022. DOI: 10.1016/j.eswa.2021.115962. URL https:// doi.org/10.1016/j.eswa.2021.115962.

[31] L. Rutkowski, M. Jaworski, and P. Duda.
Basic concepts of data stream mining. In *Studies in Big Data*, pages 13–33. Springer International Publishing, Mar. 2019. DOI: 10.1007/978-3-030-13962-9 2. URL https:// doi.org/10.1007/978-3-030-13962-9 2.

[32] Y. Spenrath and M. Hassani. Predicting business process bottlenecks in online events streams under concept drifts. In *ECMS 2020 Proceedings*

PUC-Rio - Certificação Digital Nº 2020862/CA

edited by Mike Steglich, Christian Mueller, Gaby Neumann, Mathias Walther. ECMS, June 2020. DOI: 10.7148/2020-0190. URL https://doi.org/10.7148/2020-0190.

[33] L. Su and Z. Xiao. Testing structural change in time-series nonparametric regression models. *Statistics and its Interface*, 1(2):347, 2008.

[34] F. Taymouri, M. la Rosa, M. Dumas, and F. M. Maggi. Business process variant analysis: Survey and classification. *CoRR*, abs/1911.07582, 2019.

[35] W. van der Aalst. Process mining. *Communications of the ACM*, 55(8):76–83, Aug. 2012. DOI: 10.1145/2240236.2240257. URL https://doi.org/10.1145/2240236.2240257.

[36] W. van der Aalst, A. Adriansyah, A. K. A. de Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van den Brand, R. Brandtjen, J. Buijs, A. Burattin, J. Carmona, M. Castellanos, J. Claes, J. Cook, N. Costantini, F. Curbera, E. Damiani, M. de Leoni, P. Delias, B. F. van Dongen, M. Dumas, S. Dustdar, D. Fahland, D. R. Ferreira, W. Gaaloul, F. van Geffen, S. Goel, C. Gu¨nther, A. Guzzo, P. Harmon, A. ter Hofstede, J. Hoogland, J. E. Ingvaldsen, K. Kato, R. Kuhn, A. Kumar, M. L. Rosa, F. Maggi, D. Malerba, R. S. Mans, A. Manuel, M. McCreesh, P. Mello, J. Mendling, M. Montali, H. R. MotahariNezhad, M. zur Muehlen, J. Munoz-Gama, L. Pontieri, J. Ribeiro, A. Rozinat, H. S. P´erez, R. S. P´erez, M. Sepu´lveda, J. Sinur, P. Soffer, M. Song, A. Sperduti, G. Stilo, C. Stoel, K. Swenson, M. Talamo, W. Tan, C. Turner, J. Vanthienen, G. Varvaressos, E. Verbeek, M. Verdonk, R. Vigo, J. Wang, B. Weber, M. Weidlich, T. Weijters, L. Wen, M. Westergaard, and M. Wynn. Process mining manifesto. In *Business Process Management Workshops*, pages 169–194. Springer Berlin Heidelberg, 2012. DOI: 10.1007/ 978-3-642-28108-2 19. URL https://doi.org/ 10.1007/978-3-642-28108-2 19.

[37] A. Van Looy and A. Shafagatova. Business process performance measurement: a structured literature review of indicators, measures and metrics. *SpringerPlus*, 5(1):1–24, 2016.

[38] S. J. van Zelst, B. F. van Dongen, and W. M. P. van der Aalst. Event stream-based process discovery using abstract representations. *Knowledge and Information Systems*, 54(2):407–435, May 2017. DOI: 10.1007/ s10115-017-1060-2. URL https://doi.org/10. 1007/s10115-017-1060-2.

PUC-Rio - Certificação Digital Nº 2020862/CA

[39] K. Wang, S. Zhou, and Y. He. Growing decision trees on support-less association rules. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00*. ACM Press, 2000. DOI: 10.1145/347090.347147. URL https://doi.org/10.1145/347090.347147.

[40] L. Wen, J. Wang, W. M. P. van der Aalst, B. Huang, and J. Sun. A novel approach for process mining based on event types. *Journal of Intelligent Information Systems*, 32(2):163–190, Jan. 2008. DOI: 10.1007/ s10844-007-0052-1. URL https://doi.org/10. 1007/s10844-007-0052-1.

[41] S. C. Yusta. Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters*, 30(5):525–534, Apr. 2009. DOI: 10.1016/j.patrec.2008.11. 012. URL https://doi.org/10.1016/j.patrec. 2008.11.012.

[42] R. Zaman, M. Hassani, and B. F. van Dongen. A framework for efficient memory utilization in online conformance checking, 2021. URL https://arxiv.org/abs/2112.13640.

[43] A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. strucchange: An R Package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2), 2002. DOI: 10.18637/jss.v007.i02. URL https://doi.org/10.18637/jss.v007.i02.

[44] A. Zeileis, A. Shah, and I. Patnaik. Testing, monitoring, and dating structural changes in exchange rate regimes. *Computational Statistics & Data Analysis*, 54(6):1696–1706, June 2010. DOI: 10.1016/j.csda.2009.12.005. URL https://doi.org/10.1016/j.csda.2009.12.005.

[45] FUHRMAN, J. D. et al. A review of explainable and interpretable AI with applications in COVID-19 imaging. Medical Physics, v. 49, n. 1, p. 1–14, 7 Dez. 2021.

[46] AGGARWAL, C. C. Towards effective and interpretable data mining by visual interaction. ACM SIGKDD Explorations Newsletter, v. 3, n. 2, p. 11, 1 Jan. 2002.

[47] COUSSEMENT, K.; BENOIT, D. F. Interpretable data science for decision making. Decision Support Systems, p. 113664, Ago. 2021.

PUC-Rio - Certificação Digital Nº 2020862/CA

[48] WANG, F.; KAUSHAL, R.; KHULLAR, D. Should Health Care Demand Interpretable Artificial Intelligence or Accept “Black Box” Medicine?. Annals of Internal Medicine, v. 172, n. 1, p. 59, 17 dez. 2019.

[49] SHAHBAZ, M. et al. Investigating the adoption of big data analytics in healthcare: the moderating role of resistance to change. Journal of Big Data, v. 6, n. 1, 31 jan. 2019.

[50] FABRIZIO, M.; MAGGI. Declarative Process Mining with the Declare Component of ProM. BPM (Demos) [s.l: s.n.]. Available at: <http://ceur-ws.org/Vol-1021/paper_8.pdf>.

[51] GROHMANN, N.; VOSSEN, G. Data Science Methods for Declarative Process Mining. [s.l: s.n.]. Available at: <http://ceur-ws.org/Vol-2867/paper2.pdf>. Acesso em: 14 jul. 2022.

[52] CHINOSI, M.; TROMBETTA, A. BPMN: An introduction to the standard. Computer Standards & Interfaces, v. 34, n. 1, p. 124–134, Jan. 2012.

[53] JUST, M. A.; CARPENTER, P. A. Comprehension of negation with quantification. Journal of Verbal Learning and Verbal Behavior, v. 10, n. 3, p. 244–253, jun. 1971.

[54] HORN, L. R.; WANSING, H. "Negation", The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), ZALTA, E. N. (ed.), forthcoming URL = <https://plato.stanford.edu/archives/fall2022/entries/negation/>.

[55] MONTALI, M. et al. Monitoring business constraints with the event calculus. ACM Transactions on Intelligent Systems and Technology, v. 5, n. 1, p. 1–30, 1 Dez. 2013.

PUC-Rio - Certificação Digital Nº 2020862/CA