Causal and Interpretable Rules for Time Series Analysis

Amin Dhaou^{*†} TotalEnergies Palaiseau, France amin.dhaou@totalenergies.com

Antoine Bertoncello TotalEnergies Palaiseau, France antoine.bertoncello@totalenergies.com sebastien.gourvenec@totalenergies.com

Sébastien Gourvénec TotalEnergies Palaiseau, France

Josselin Garnier CMAP, Ecole Polytechnique, Institut Polytechnique de Paris Palaiseau, France josselin.garnier@polytechnique.edu

Erwan Le Pennec CMAP, Ecole Polytechnique, Institut Polytechnique de Paris Palaiseau, France erwan.le-pennec@polytechnique.edu

ABSTRACT

The number of complex infrastructures in an industrial setting is growing and is not immune to unexplained recurring events such as breakdowns or failure that can have an economic and environmental impact. To understand these phenomena, sensors have been placed on the different infrastructures to track, monitor, and control the dynamics of the systems. The causal study of these data allows predictive and prescriptive maintenance to be carried out. It helps to understand the appearance of a problem and find counterfactual outcomes to better operate and defuse the event.

In this paper, we introduce a novel approach combining the case-crossover design which is used to investigate acute triggers of diseases in epidemiology, and the Apriori algorithm which is a data mining technique allowing to find relevant rules in a dataset. The resulting time series causal algorithm extracts interesting rules in our application case which is a non-linear time series dataset. In addition, a predictive rule-based algorithm demonstrates the potential of the proposed method.

CCS CONCEPTS

• Information systems \rightarrow Association rules; Data mining; • Computing methodologies \rightarrow Supervised learning by classification; Rule learning; • Applied computing \rightarrow Command and control.

KEYWORDS

Causality, Time Series, Data Mining, Case-Crossover design, Predictive maintenance

*Corresponding author

KDD '21, August 14-18, 2021, Virtual Event, Singapore

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-8332-5/21/08...\$15.00 https://doi.org/10.1145/3447548.3467161

ACM Reference Format:

Amin Dhaou, Antoine Bertoncello, Sébastien Gourvénec, Josselin Garnier, and Erwan Le Pennec. 2021. Causal and Interpretable Rules for Time Series Analysis. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14-18, 2021, Virtual Event, Singapore. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548. 3467161

1 INTRODUCTION

Monitoring has enabled, with the help of increased storage capacity, to collect a large amount of data. The data analysis plays a crucial role in understanding the underlying mechanisms and the occurrence of incidents. In the industrial context, this consists of placing sensors and collecting temporal data like temperature, flow rates, chemical characteristics, or wind power to capture the evolution and the dynamics of the system. Exploiting these large amounts of temporal data is a real challenge facing many companies. Indeed, they contain enormous amounts of information that could help improve efficiency or optimize certain processes.

Driven by easy access to machine learning environments and the recent success of deep learning techniques, many models have been developed to predict the occurrence of these events but they do not only work on their causes but also on the correlated variables. This makes these models less robust as they could miss the incident by trusting a correlated variable. In areas where decisions and actions can have serious consequences, for example on humans in medicine or on the profitability in the industry, it is necessary to understand black-box models and therefore to carry out a causal study to act in a justified way. Hence, the objective of causality in an industrial context is to better understand the decisions taken by artificial intelligence algorithms, to find the causes of unexplained events, and to do maintenance policy by anticipating the occurrences of breakdowns. Therefore, a theoretical approach should be developed to provide a general framework that could work in an industrial environment. In particular, the approach should help the operators understand what are the mechanisms behind every decision that is taken and allow them to prevent the apparition of an incident by defusing its arrival.

The interest in causality is growing and these studies are becoming essential in industry and in many other fields of applications. For instance, it is common for distillation units to have recurrent problems occurring during petroleum refining. The causal study

[†]Also with CMAP, Ecole Polytechnique, Institut Polytechnique de Paris.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

allows a better understanding of the origins of these problems and to develop a general approach that can be used on many systems such as wind turbines.

Petroleum refining is the process of transforming crude oil into usable end products such as gasoline, diesel, and other products used in petrochemicals. The distillation process is the first stage of processing and consists of purifying different liquid substances from a mixture. In our case, this is done by separating the different hydrocarbon fractions contained in crude oil. Regularly, an event called flooding [18, 19] occurs and requires the process to be slowed down or stopped for a considerable amount of time. In fact, several hours of maintenance are required to return to a normal operating regime. This happens when the steam flow is too high and blocks the flow of liquid in the column. It could be detected by sharp increases in differential pressures, and a decrease in the production performance. In fact, a flooding results in a loss of performance and a decrease in the quality of separation. In the case of the atmospheric distillation unit, a flooding occurs very often and several hours are then required to stabilize the unit.

Flooding events are complex and the origins are still unclear but we know that there are of different types. Conventional methods based on theoretical equations and/or temperature and differential pressure analysis have been used to develop predictors. These attempts have so far been unsatisfactory; either the number of False Positives is too high, or a large number of floodings are missed. Currently, a predictive model has been established in order to predict the onset of the floodings. Indeed, the use of a Random Forest algorithm allows to predict the problem one hour before its appearance but even though the model achieves good accuracy, there are still many False Positives and some events are not detected. This leads to a waste of time and an economic loss.

Two limitations of this approach can be pointed out: firstly the random forest algorithm is a black-box model which means that even if we are able to forecast the onset of a flooding, the causes remain unclear. Indeed, a random forest is a set of decision trees where each one takes a decision based on different samples and the final decision is a vote over the decisions of the trees. Secondly, it is not a causal model, hence the reliability of this algorithm is reduced.

Once the predictive model has been implemented on-site to predict in real-time the apparition of the floodings, the operators have asked several questions: how does the algorithm make its predictions? Why should they trust a model that they do not understand and cannot be explained? What should be done to prevent the flooding when the alarm is triggered? These questions have remained unanswered and we believe it is essential to find answers. A causal study is thus necessary to give a more explainable relationship between the variables and the onset of the event.

Causality has always been a subject of much research in both science and philosophy. From an observable effect, we want to understand what are the causes of this phenomenon. The mechanism of cause and effect can be visible (domino) but it can be much more complex with unobservable steps, as in chemical processes. Therefore, links can be established between different independent or dependent variables jointly called correlations, and often confused with causality although it is different. Several definitions of causality and their associated approaches have been developed in order to extract the causal structure from the data, of which the best known are Granger causality [14] and Pearl causality [26].

Our study focuses on time series, so we need to establish a suitable framework. Granger proposed a "predictive causality" test to determine whether one time series improves the predictive ability of a second. This approach is limited because hidden variables could reveal correlations rather than causality. On the other hand, Pearl proposes the "Structural Causal Model" allowing to establish a causal graph from the data. The temporal dimension corresponds to additional constraints on the structure of the graph.

Our approach, which can be described as a Granger method, is an adaptation of the case-crossover design [20] (developed in section 3) to industrial data. It is an approach used in epidemiology in order to understand the origins of a phenomenon appearing suddenly (heart attack, accidents, injuries [9-11, 22]). A relation between an exposure and an event is said to be causal if the occurrence of the exposure causes the event. Finding the causes of acute events has always been a challenge for epidemiologists and the way the data collected from a batch of patients is analyzed plays an important and crucial role in the study.

This design is combined with the association rule mining algorithm Apriori [27] which aims at discovering relationships of interest between two or more variables stored in datasets. The advantage of this method is that it has a high interpretability [34], hence easier to understand for operators that could then be able to act and defuse the problem.

In this paper, we present a general framework that could be used when we want to understand the causes of a phenomenon that occurs briefly over time and we demonstrate its relevance by applying it to our flooding case study. We propose the Case-crossover APriori (CAP) algorithm which provides association and causal rules explaining the occurrences of flooding events and the Casecrossover APriori Predictive algorithms (CAPP1 and CAPP2) that predict them. The purpose of this work is to answer the following research questions: What are the causes of the flooding event and what are the variables involved in this phenomenon? The paper is organized as follows. Section 2 is dedicated to a review of existing literature on causality for time series, the interpretability of algorithms and causal rule mining. Section 3 is the presentation of the case-crossover design. Section 4 contains a depiction of the original methodology developed in this study. Section 5 is dedicated to the case study used to validate the research approach with the presentation of the data and the results. Section 6 contains the conclusive remarks and approaches for future work.

2 RELATED WORK

Temporal data is complex and surrounds us in everyday life. Their study in industrial and mechanical environments is fundamental, it increases prevention and reduces the occurrence of problematic and damaging phenomena.

Causality being a rather old field, a lot of research has been done and several approaches exist. Many of these approaches have been adapted to temporal data adding order constraints on the data for each variable. In the case of Pearl causality and graphical approaches based on Structural Equation Models, we can cite the algorithms PCMCI [31] and tsFCI [8] which are, respectively, adaptations of the non-temporal causality algorithms PC [33] and FCI [33]. The general principle is based on two steps. The first step is about statistical testing to establish the conditional dependencies between the variables and thus obtain the skeleton of the graph. The second step then constrains the types of existing relationships and orients the edges of the graph. The addition of the temporal dimension has the effect of adding an assumption on the precedence associated to each node of the graph. Indeed, future elements cannot be the causes of past elements.

The approaches proposed by Granger have been also adapted to time series. In fact, the reformulation of the "Granger causality" is as follows: it is said that X "Granger causes" Y if the future values of Y (at time t + 1) can be better predicted by using the previous values of X and Y (up to time t) than by using only the previous values of Y. Many papers have used this approach with notably vector autoregressive models [6] and kernel-based methods [23]. Predictive models are then followed by a statistical test on the residuals which can be of several forms: SSR-based F test or Pearson Chi-square test.

In addition, several approaches that can be related to Granger causality have been developed to explain machine learning models, also called black-box models because they are very difficult to interpret due to the high number of parameters involved and the complexity of these models. More recently, different approaches based on Granger causality have been developed such as copulabased methods [16], regression techniques [12] or Deep Learning method with Attention model [25].

Other types of methods to explain the decisions of black-box models have been developed. In particular, we can mention LIME [30] and LoRMIkA [28] which use association rules. These methods allow to explain locally the decisions made by the algorithm but suffer from a lack of interpretability at the global level.

The first method that could be compared with our algorithm is decision trees. This is a very popular approach as it is simple to compute and highly interpretable, we can cite the two main algorithms CART [4] and C5.0 [32]. The drawbacks are that they are very unstable and are not good predictors. That is why Breiman developed random forest [3] to overcome these issues but the complexity introduced by the high number of trees and the bagging technique reduces the interpretability. On the other hand, the second method is the rule mining algorithm in which we can find greedy heuristics [7, 13, 24, 35] allowing us to find interpretable rules. Similar to decision trees, they have low accuracy and are unstable. More recently, Benard et al. [5] extracted interpretable rules from a random forest classifier by looking for frequent patterns in the trees, hence allowing to take advantage of the high accuracy and the stability of the algorithm, while being interpretable.

We should note that these methods have been developed for non-temporal data. Although the random forest algorithm can be used for temporal data, it does not take into account trend and seasonality since it only gives thresholds based on a criterion that separates the classes so that we have two homogeneous groups. Likewise, the rule-based methods must be adapted to respect the temporal aspect and to take into account the trends and seasonality.

Our approach overcomes these problems by using the casecrossover design, which has been developed for time series data. Then, the use of the Apriori algorithm allows having interpretable and causal rules.

3 CASE-CROSSOVER DESIGN

The method used in this paper is inspired by an epidemiological approach and we aim here at giving an intuitive introduction. The most natural way to find the cause of a disease that happens briefly across a population is to take two subgroups, one that contracted the disease and the other one that is healthy, and then compare them. This design called the case-control design must fulfill some conditions to remove biases from the study. In fact, the two subgroups that we compare should have the same characteristics like age or gender in order to remove confusion bias called confoundings. They could create an association that does not exist or hide an existing one and this design is not tailored to prevent or control them.

The case-crossover design, proposed by Maclure in 1991 [4], is used in epidemiology to study the onset of acute events across a population and is widely used to find the causes of diseases. It is an alternative to the case-control design and allows to avoid confusion biases. There are many applications of this design that are used such as studying the effect of air pollution on health [5] or more recently looking for the causes of the increase of mortality due to Covid-19 and cold temperature [6]. The idea behind this design is to take two periods, one named the control and the other one the case period. The control period is selected in a "normal" operating period i.e. a long time before the onset of the event and the case period is selected during the hazard period i.e. the period preceding the onset of the event. The comparison of these two periods tells us what has changed between these two periods and statistics over several events allow us to see which changes happen regularly across a population and induce the event or the disease. Let us take the example of a car accident [21] to understand why this design is relevant. To know what the causes of car accidents are in an area, the case-control design would identify two groups of people. The first group is made up of the people who had an accident in that area - the case group - and the second group is made up of the people who did not have an accident in the same area - the control group. Then, we compare their behaviors to see if there is a factor that happens often in the "case group" and not in the "control group". For example, we could see that in the case group, people are more likely to eat while driving or to have a phone call, and then we could identify these behaviors as a cause. This approach has some issues as we compare different people that have different characteristics. If we do not preprocess our data and do not compare similar subjects, it introduces confoundings.

The case-crossover design aims at tackling this issue. In fact, instead of taking two groups, we only follow the subjects who had an accident. The control period would be a long time before the accident and the case period the moments just before the event. For the car accident example, it allows us to see that the subject was not using his phone during the control period and was using it during the case one. By doing these comparisons of the same subject but at different time periods, we could remove the confoundings coming from the differences between different subjects taken as case and control. Another advantage is that we do not need the second group anymore but only subjects who are exposed.

Usually, this process is done through several individuals taken on the same day of the week and at the same time over several months. The results are then given by doing statistics on the overall data collected. We insist here on the fact that it is necessary to have prior knowledge of the duration of exposure and the appearance of the phenomenon.

4 RULE-BASED ALGORITHM

Association Rule Mining (ARM) is a data mining framework that allows to extract frequent associations of variables in a database. It has the advantage of being highly interpretative and easy to understand. In this section, we describe how ARM has been used in retail and how we adapt it to more varied fields of application and particularly for time series by introducing the CAP, CAPP1 and CAPP2 algorithms.

4.1 Motivation

ARM has been developed by Agrawal [17] for commercial purposes. Indeed, commercial enterprises accumulate a significant amount of data on a daily basis. In the case of supermarkets, consumer purchases that can be retrieved from checkout receipts are a huge source of information. Their analysis helps to better understand consumers' behavior and thus to establish appropriate marketing campaigns, better manage inventories or improve customer relations.

The general setting for ARM is composed of a database containing transactions and each transaction is an item-set i.e. a set of items. ARM algorithms allow to find relationships between items from the database in the form of association rules which are rules (implications) of the form $a \rightarrow b$ where *a* is an item-set and *b* is an item-set that is not present in *a*. In our case, we consider that *b* is only one item.

4.2 Apriori

In the following, we use the Apriori algorithm. Our objective being to extract causal rules from a database, we consider the use of this algorithm sufficient. An improvement of the algorithm will be made in future works.

Several problems can arise when using association rules. The number of rules depends on the size of the database, if it is large, it will be impossible to consider all the possible rules. In fact, in a database with *n* items, the number of rules of the form $a \rightarrow b$ for all possible item-sets *a* and items *b* that are not present in *a* is $n2^{n-1}$, hence the complexity would be exponential and the problem intractable. Moreover, we will find among all these rules random associations and the items coming into play would have no real causal link or association.

The paradigm usually applied for mining rules is the Frequent Rule Mining (FRM). In order to proceed, we need to define the framework to determine the association rules. Let *I* be the set of items and *D* be the set of transactions, which is a set of item-sets of *I*, and let *a* be an item-set and *b* an item. Rules extracted by FRM are of the form $a \rightarrow b$. In the following, we give definitions and metrics [2] in order to apply the algorithm.

• The **support** supp $(a \rightarrow b)$ ranges within [0, 1] and is the frequency of apparition of *a* and *b* within the transactions. Here, $\#(a \cup b)$ is the number of elements of *D* that contain the item-set $a \cup b$.

$$\operatorname{supp} (a \to b) = \frac{\#(a \cup b)}{|D|}$$

The rules extracted by FRM have their support greater than a set threshold called minimum-support.

We also define supp(a), where #(a) is the number of elements of *D* that contains *a*

supp
$$(a) = \frac{\#(a)}{|D|}$$

• The **confidence** ranges within [0, 1] and is the percentage of transactions containing *a* that also contain *b*. It is an estimate of the probability of observing *b* given *a* and is an indication of how often the rule has proven to be true.

$$\operatorname{conf} (a \to b) = \frac{\operatorname{supp} (a \to b)}{\operatorname{supp}(a)}$$

Confidence is directed which means that the confidence of the rules $a \rightarrow b$ and $b \rightarrow a$ are different.

• The **lift** ranges withing $[0, +\infty]$ and is the ratio of the observed support to that expected if *a* and *b* were independent and indicates the strength of an association rule over the random occurrence of *a* and *b* in a transaction.

lift
$$(a \rightarrow b) = \frac{\text{supp } (a \rightarrow b)}{\text{supp}(a) \times \text{supp}(b)}$$

In fact, rules with high confidence can occur by chance. That is why we need to add a measure of independence to find these spurious associations. If the lift is close to 1, it means that a and b are independent and the rule is not interesting whereas if the lift is large or close to 0 it means that a and bare associated. The lift is symmetric and is not able to capture the rule direction.

The conviction ranges within [0,+∞] and can be interpreted as the ratio of the expected frequency that *a* occurs without *b*, that is to say, the frequency that the rule makes an incorrect prediction.

$$\operatorname{conv}(a \to b) = \frac{1 - \operatorname{supp}(b)}{\operatorname{conf}(a \to b)}$$

The conviction gives additional information to confidence and lift as it gives a metric for the notion of implication in the rule. High conviction values mean that the rule is interesting.

4.3 Notations

In the following, we focus on the supervised binary classification framework using the case-crossover design. In the section 5, we show that an appropriate preprocessing of the dataset allows the selection of control and case periods on the same time series ("individual"). Let us denote for each time series $(Z_t)_{t=0,...,T}$,

for each $(\tau_1, \tau_2) \in \{0, ..., T\}^2$ such that $\tau_1 < \tau_2, Z_{[\tau_1, \tau_2]} = (Z_t)_{t=\tau_1,...,\tau_2}$, hence for $\tau \in \{0, ..., T\}, Z_{[0,\tau-1]} = (Z_t)_{t=0,...,\tau-1}$ contains the past values of Z_{τ} . Suppose, that we have a sample composed of *n* pairs $\mathcal{D}_n = \{((X_{it})_{t=0,...,T}, Y_i), i = 1,...,n\}$ where the *n*



Figure 1: Proposed design applied on a time series without the event: control

pairs are i.i.d of law $((X_t)_{t=0,...,T}, Y)$. For all $t, X_t = (X_t^{(1)}, \ldots, X_t^{(p)})$ is a random vector taking values in \mathbb{R}^p and $Y \in \{0, 1\}$ is the binary outcome. For a time series $(X_t)_{t=0,...,T}$, the goal is to predict the binary output *Y*. Hence, the objective is to find an interpretable and causal predictive model of the event Y = 1 given $(X_t)_{t=0,...,T}$.

In this section, we propose an original method inspired by the case-control and the case-crossover designs which can process continuous or categorical temporal data. The method aims at finding an interpretable and causal predictive model of the event Y = 1 given $(X_{[0,\delta]}, X_{[T-\delta,T]})$ where δ is the duration of a period and $\Delta = T - 2\delta$ is the gap between the two periods, as shown in Figures 1 and 2. First, using prior and domain knowledge, we select the periods allowing us to characterize the event. Secondly, we need to transform the periods into a categorical dataset. Since association rule mining algorithm only works with categorical data, we should indeed apply a transformation to convert continuous variables into categorical data without losing relevant information. This allows to extract simple rules explaining the dynamics of the phenomenon.

4.4 Methodology

We decide to apply the case-crossover design on a dataset using association rule mining by creating an algorithm called Case-crossover APriori (**CAP**). The first step is to set up an environment in which we are able to compute rules. We need to define what our "transactions" are and the type of "items" that will be included in our rules. As rules are computed between periods of the time series, we need to set a metric that creates the items.

Case-crossover design. We first need to adapt the case-crossover design to a time series dataset designed for classification. It is done by constructing a parametric model that could be optimized allowing the user to select the control and case periods, to fine-tune the model by selecting the best parameters.

In our dataset, we have case samples where we have the event $(Y_i = 1)$ and control samples where we don't $(Y_i = 0)$. This approach allows us to use a machine learning framework for binary classification whether we want to compute interpretable rules or do predictions. We decided to define periods as shown in Figure 1 and 2.

In this design, we are doing comparisons between period₁ and period₂ that have the same duration δ (which is an hyper-parameter) and that are separated by an interval of duration $\Delta = T - 2\delta$. Hence, for each time series $(X_t)_{t=0,...,T}$ taken in \mathcal{D}_n , we can extract the first period $X_{\text{period}_1} = X_{[0,\delta]}$ and the second period $X_{\text{period}_2} = X_{[T-\delta,T]}$.



Figure 2: Proposed design applied on a time series with the event: case

We explored different ways to compare the periods without losing the dynamic of the system. We decided for each variable to map the data of each period into one value using different transformations. For each period of duration δ , we calculated the standard deviation, the mean, and the residuals of autoregressive models for each variable. The most convenient and relevant way we found is to compute the mean of each selected period for each variable. This transformation is easy to understand and interpretable and we could find a counterfactual based on it.

we could find a counterfactual based on it. Once we have the values $\bar{x}_{\text{period}_1} = (\bar{x}_{\text{period}_1}^{(j)})_{j=1}^p$ and $\bar{x}_{\text{period}_2} = (\bar{x}_{\text{period}_2}^{(j)})_{j=1}^p$ summing up the dynamics of the system during the first and second periods, we need to compare them. The metric chosen is the **percentage change**:

$$\left(\bar{x}_{\text{period}_1}^{(j)}, \bar{x}_{\text{period}_2}^{(j)}\right) : \to \left| \frac{Max(\bar{x}_{\text{period}_1}^{(j)}, \bar{x}_{\text{period}_2}^{(j)}) - Min(\bar{x}_{\text{period}_1}^{(j)}, \bar{x}_{\text{period}_2}^{(j)})}{Max(\bar{x}_{\text{period}_1}^{(j)}, \bar{x}_{\text{period}_2}^{(j)})} \right|$$
(1)

Let us denote by f the function taking as input the first and the second period X_{period_1} and X_{period_2} , computing the mean of each period $\bar{x}_{\text{period}_1}$ and $\bar{x}_{\text{period}_2}$ and then the percentage change using the metric (1). Finally, the problem is formulated as follows: the objective is to find a predictive model of the event

$$Y = 1$$
 given $f(X_{\text{period}_1}, X_{\text{period}_2})$

If we take the example of Figure 2, we first select the two periods shown in red in the figure. Then, we compute the mean of each period. Finally, we compute the percentage change of the means. In Figure 2, there is a large increase of the mean value between the two periods, hence a large value of the percentage change, while in Figure 1 there is no meaningful change between the two periods.

Association Rule Mining. Association rule mining algorithms, like the Apriori algorithm, take as input categorical variables. Hence, we need to do a "categorization step" because when we compare the pair $\bar{x}_{\text{period}_1}^{(j)}$ and $\bar{x}_{\text{period}_2}^{(j)}$ using the percentage change metric (1), we have values between 0 and 1 if we consider that all variables takes positive values (this hypothesis is not restrictive for continuous real-valued random variables as it is always possible to transform them into random variables with uniform distribution over [0,1]). Then, we categorize these values into three categories LOW, MEDIUM, and HIGH. This was decided for clarity reasons to explain the method, but it can be extended to an arbitrary number of categories.

Let us note for each $j \in \{1, ..., p\} X_{\{\alpha, \beta\}}^{(j)}$ the boolean variable which is True if the percentage change of $X^{(j)} = (X_t^{(j)})_{t=0,...,T}$ falls in the interval defined by the quantiles of order α and β . We first estimate two empirical quantiles from the dataset, the quantiles of order 0.33 and 0.66. Thus for each $j \in \{1, ..., p\}$, we have three boolean variables to indicate the range in which the percentage change is: $X_{\{0,0.33\}}^{(j)}, X_{\{0.33,0.66\}}^{(j)}$ and $X_{\{0.66,1\}}^{(j)}$. For each of the time series $(X_{it})_{t=0,...,T}$ taken in \mathcal{D}_n , we select

For each of the time series $(X_{it})_{t=0,...,T}$ taken in \mathcal{D}_n , we select period₁ and period₂, compute their means and compare them using the percentage change metric defined in (1). After completing the process for all the samples in the database \mathcal{D}_n , we set up the Table 1.

Event ID	Items
Y1	$(X_1^{(1)})_{\{0,0.33\}} = False, (X_1^{(1)})_{\{0.33,0.66\}} = False,$
	$(X_1^{(1)})_{\{0.66,1\}} = True, (X_1^{(2)})_{\{0,0.33\}} = True,$
	$(X_1^{(2)})_{\{0.33,0.66\}} = False, (X_1^{(2)})_{\{0.66,1\}} = False, \dots$
Y2	$\dots, (X_2^{(3)})_{\{0.33, 0.66\}} = True, \dots,$
	$(X_2^{(5)})_{\{0,0.33\}} = True, (X_2^{(5)})_{\{0.33,0.66\}} = False, \dots$
Y3	$(X_{3}^{(1)})_{\{0,0.33\}} = False, (X_{3}^{(1)})_{[0.33,66]} = False,$
	$(X_3^{(1)})_{\{0.66,1\}} = True,$
	,
Y_{n-2}	$\dots, (X_{n-2}^{(2)})_{\{0,0.33\}} = True, \dots, (X_{n-2}^{(4)})_{\{0.33,0.66\}} = True, \dots$
Y_{n-1}	$\ldots, (X_{n-1}^{(2)})_{\{0,0.33\}} = True, \ldots, (X_{n-1}^{(4)})_{\{0.33,0.66\}} = True$
Yn	$(\dots, (X_n^{(1)})_{\{0.33, 0.66\}} = False, (X_n^{(1)})_{\{0.66, 1\}} = True,$
	$(\dots, (X_n^{(3)})_{\{0,0.33\}} = True,\dots$

 Table 1: Table constructed from the comparisons of the selected periods

In the "Event ID" column, for $i \in \{1, ..., n\}$ Y_i is the outcome of the time step T + 1. If the event happens, then $Y_i = 1$, otherwise $Y_i = 0$. Hence, the comparison between the case and control allows to identify variations characteristic of the event and to separate them from independent variations. The "Items" column gathers, for each of the "Event ID", the interval in which the percentage change of each variable is.

Then we compute one-hot encoding [1] and we add a boolean column "Event" by transforming the "Event ID" column into a label. We then apply the Apriori algorithm and can get the rules leading to an event.

We used the Apriori algorithm from the package mlxtend [29]. In this library, the user should specify the minimum support with the parameter **min_support** in order to find frequent item-sets. Thresholds on the metrics **confidence**, **lift** and **conviction** defined above could also be fine-tuned if we want to further discriminate the rules. Moreover, the number of items and associations in the rule can be set using the parameter **max_len**. Finally, by adding a constraint to have only rules which have a target "Event=True" i.e. Y = 1, and a max_len of 1, the CAP algorithm could find rules like:

$$[X_{\{0,0.33\}}^{(1)} = True\} \implies \{\text{Event}=\text{True}\}$$

4.5 **Predictive Algorithm**

Beyond the evaluation of the rules found by the Apriori algorithm which is made by experts, we want to test predictive properties by creating a first predictive algorithm that we call Case-crossover APriori Predictive 1 (**CAPP1**). The goal is to predict the binary output *Y* based on simple and understandable rules. We selected the first 10 rules by order of confidence and lift to do the prediction on a test time series *X* of length *T*.

Event ID	$(X^{(1)})_{\{0,0.33\}}$	$(X^{(1)})_{\{0.33, 0.66\}}$	$(X^{(1)})_{\{0.66,1\}}$	$(X^{(2)})_{\{0,0.33\}}$	$(X^{(2)})_{\{0.33, 0.66\}}$	
Y_1	0	0	1	1	0	
Y2	0	1	0	0	0	
Y3	0	0	1	0	0	
Y_4	1	0	0	0	1	
Y ₅	0	1	0	1	0	

Table 2: One-hot encoding

In Figure 2, we compute the left-hand side of the implication symbol of the 10 rules that have been found. If at least one rule is True, CAPP1 predicts an "event" and triggers the alarm. We have experimented several other approaches to perform rule-based prediction, among them we can cite the simple aggregation technique which consists of voting on the found rules similar to what the SIRUS algorithm does [5]. The aggregation could be improved by using an ensemble learning method such as stacking [15] by learning the decision combining these 10 rules. The perfect predictive algorithm would predict an "event" for the Figure 2 but not for the Figure 1.

Example. To better understand the process, let us take the example using the first rule of Table 3.

$$\{X_{\{0,0.33\}}^{(1)} = True, X_{\{0.33,0.66\}}^{(2)} = False\} \implies \{\text{Event=True}\}$$

We consider a test time series $(X_t)_{t=0,...,T}$ and select the first and second variables $X^{(1)}$ and $X^{(2)}$ and compute their percentage changes between period₁ and period₂. If the percentage change of the variable $X^{(1)}$ is less than the quantile 0.33 and that of the variable $X^{(2)}$ is not between the quantiles 0.33 and 0.66, the algorithm predicts an event.

In order to estimate the error of our predictive model, we need to classify the predictions into four outcomes: the True Positive (TP), the True Negative (TN), the False Positive (FP), and the False Negative (FN). Then, we use the following metrics:

• True Positive Rate (TPR) or Recall summarizes the fraction of examples assigned to the positive class that belongs to the positive class

$$TPR = \frac{TP}{TP + FN}$$

• Similarly, True Negative Rate (TNR) summarizes how well the negative class is predicted

$$TNR = \frac{TN}{TN + FP}$$



Figure 3: This figure shows how from one time series of duration 20 hours (1200 minutes), we make a cutout to obtain the control sample in green and the case sample in red.

• F2-score is a weighted F-score and used when it is much worse to miss a True Positive than giving a False Positive

$$(1+2^2) \times \frac{precision \times recall}{(2^2 \times precision) + recall}$$

These metrics can be used on the training database, on a test set, and in cross-validation.

Case-crossover APriori Predictive 2 (CAPP2). A complementary approach called Case-crossover APriori Predictive 2 (CAPP2) has been studied in order to improve the quality of prediction. Indeed, in addition to looking for the rules leading to an event of the form

$$\{X^{(1)}_{\{0,0.33\}} = True, X^{(2)}_{\{0.66,1\}} = True\} \implies \{\text{Event} = \text{True}\}$$

we have also looked for the contraposed, which are the rules that do not lead to an event (leading to "Event = False") of the form:

 $\{X^{(4)}_{\{0.33,0.66\}} = True, X^{(3)}_{\{0.33,0.66\}} = True \} \implies \{\text{Event} = \text{False}\}$ Let us call "Event=True rules" the first rules and "Event=False

Let us call "Event=True rules" the first rules and "Event=False rules" the second ones. There are different ways to combine these two approaches in order to compute a more robust predictive model. Among them, we could adjust the number of rules proving to be True for each of the two approaches, we could give more weight to the "Event=True rules" for the prediction or give more weight to the "Event=False rules". Cross-validation allows to test, observe and study the behavior of each of these experiments. The decision could be improved by learning the decision combining the two types of rules.

5 APPLICATION

5.1 Data

Numerous sensors were placed at various points in the distillation unit to collect data and monitor the evolution of the system. More than 800 variables were measured, providing information such as the type of input crude, pressures, temperatures, flow rates, valve openings, and chemical measurements. The variables are categorical or continuous and take positive values. These measurements were carried out every minute for 4 months and the identification of flooding events is calculated using a formula involving variables from the outputs of the distillation column and is presented in the data in the form of an additional binary column where a 0 represents a normal operation condition system and 1 represents the flooding event. Each column represents a measured variable and each line describes the system at a specific minute.

In our study, we consider that flooding events are independent of each other. For this reason, we only take into account events that occur at least 20 hours apart from each other. Thus, we identify a total of 38 long time series (of duration 20 hours=1200 minutes) to be studied. We have therefore cut the data into 38 long time series where the last moments correspond to the appearance of the flooding event. In order to build the database \mathcal{D}_n using the case-crossover design, we must have pairs $(X_{t=0,...,T}, Y)$. Therefore, we need to define the duration T of the time series and get samples such that we have couples with labels Y = 1 and Y = 0 coming from the same long time series. The label Y = 1 is simple to obtain because we just have to select the last moments of each of the 38 time series because by definition they all end with a flooding event. For the label Y = 0, we had to sample and select a part of the 38 series. Since we assume that the samples are independent, we have to select this period so that it is far enough from the flooding event and under normal operating conditions. With the advice of experts, we decided to select samples at a time distance of 10 hours=600 minutes from the flooding event. This step of selection of periods requires preliminary knowledge of the phenomenon in order to select the periods of "normal" and "abnormal" functioning. In our case, we know that the event is acute and occurs in the hour before the event.

Figure 3 summarizes the principle of the case-crossover design and highlights the data cutout to obtain the control and case of Figure 1 and Figure 2.

Therefore, to learn rules, we have a training database $\mathcal{D}_n = \{((X_{it})_{t=0,...,T}, Y_i), i = 1, ..., n\}$ where n = 76. 38 samples of \mathcal{D}_n have a label $Y_i = 1$ and 38 samples have a label $Y_i = 0$. The sampling is done every minute and we have 4 hours of measurements for each sample, hence T = 240.

5.2 Interpretable Rules found by CAP

In this subsection, we use expert knowledge of the characteristic times of important phenomena to determine certain time parameters such as δ . For the rest of the parameters, we did not want to optimize them too much to avoid overfitting, optimizing the thresholds is an idea to keep in mind if the learning base is large enough.

After preprocessing the data, we computed the Apriori algorithm with the described design with a period duration of $\delta = 60$, 1 hour sampled every minute, and a gap $\Delta = 120$ of 2 hours between period₁ and period₂. We set *min_support* ≥ 0.2 and *min_len* = 2 and sort the results by confidence and lift. The rules that have been found are shown in Table 3.

Among the rules, we can see the presence of $X^{(1)}$ which is a variable computed from a physical model and used to be, before the random forest model, the variable allowing to determine the appearances of flooding events. Moreover, $X^{(2)}$ is a flow recirculation variable and has been selected by experts as being very likely to explain the flooding appearance.

5.3 CAPP1 Prediction Results

To prevent overfitting and evaluate well the CAPP1 method performance, we decided to do a Leave-Two-Out (LTO). For $j \in \{1, ..., n/2\}$, we take the $(2j - 1)^{th}$ and $(2j)^{th}$ element of the database \mathcal{D}_n for testing, such that we have a couple computed from one of the 38 long time series with one element having a label Y = 0 and the other Y = 1, and we take the n - 2 other elements of \mathcal{D}_n as a training set. The training set provides data to the Apriori algorithm in order to learn rules using the different metrics we defined. The rules are then sorted by confidence and lift and are ready to be tested. For the testing, as described in subsection 4.5, we predict the two elements in the test set. Finally, we evaluate the prediction by computing the True Positive Rate, the True Negative Rate, and the F2 score and compute the mean of these scores over the 38 tests we have done with our LTO. Thus, in the following, all calculated scores are obtained by cross-validation.

As mentioned in section 4.5, we select the 10 rules with highest confidence and lift, and with two or fewer explanatory variables, then we calculate the quality of the prediction using the defined metrics. We evaluate the predictive performance of the CAPP1 method by a comparison with the one of a random forest (RF) algorithm. The RF is trained with the dataset D_n and takes as input the averages of the input variables over $[T - \delta, T], [T - 2\delta, T - \delta], \ldots, [0, \delta]$ and predicts the binary label "there is a flooding at time T + 1 minute". The results are shown in Table 4.

We could always increase the True Positive Rate by choosing a higher threshold for the minimum support and increasing the number of rules but this will directly affect the True Negative Rate as there is a trade-off between True Positives and False Positives. If our model is more sensitive and often rings an alarm, it will make more errors and then more False Positives.

The results are satisfactory as the True Positive Rate is relatively high and far better than a random prediction without even optimizing our algorithm but are insufficient compared to the random forest algorithm.

5.4 CAPP2 Prediction Results

After several tests, we opted for the following combination: we set $min_support \ge 0.01$ and sort the results by confidence with a minimum threshold of 0.5. If at least one out of the first 100 "Event=True rules" and less than one out of the first 100 "Event=False rules" is True, we predict that the tested pair leads to a flooding event i.e. Y = 1. Otherwise, we predict that the pair does not lead to a flooding event i.e. Y = 0. Since a minimum threshold of confidence has been set, the number of rules can be lower but limited to 100. Note that the choice of 100 rules here is empirical and depends on the choice of the minimum support threshold.

CAPP2 has allowed us to improve our prediction results and obtain the scores presented in Table 4.

Algorithm	F2 score	TNR	TPR/recall			
Random Forest	0.8127	0.8368	0.8684			
CAPP1	0.6991	0.6644	0.8684			
CAPP2	0.9139	0.9210	0.8947			

Table 4: Prediction scores.

These results are promising as the CAPP2 method achieves better scores than RF without optimizing our model with a relatively small dataset and especially with a model that proposes a causal analysis.

6 CONCLUSION AND FUTURE WORK

We have developed a data-driven model based on the case-crossover design and association rule mining for determining the causes of an incident from time series. This approach overcomes two main issues: the lack of interpretability and prediction based on correlations. The understanding of incidents is essential because it would allow to predict in advance their appearance using a causal prediction algorithm and to be able to justify the reliability and confidence contrarily to a black-box algorithm.

The application and study of this approach to our dataset provide conclusive results confirming that the method is promising. This work gives insight to operators working in the refinery with the distillation unit and allows them to understand the mechanisms that trigger the event. The method finds interesting rules and describes associations between variables leading to an event. Among the top rules sorted by confidence, we find the variables that have been suspected to be causal by the experts. The associations make it possible to strengthen them and to add missing information necessary to the understanding of the phenomenon of flooding. In addition, our predictive study has shown that we could build a strong predictive model which could outperform the one actually in production. Indeed, the results on the four-month dataset have confirmed these expectations and there is still a lot of room for improvement.

This method uses expert knowledge to select certain parameters. In the absence of such information, methods to determine these characteristic times must be considered and more failure case data may be needed for this.

Several approaches have been identified for future work. Among them, we could cite the following ideas: instead of choosing two arbitrary quantiles as we did in this work, we could optimize them and adapt their number. We could also deepen the contraposed

rules	support	confidence	lift
$\{X_{\{0,0.33\}}^{(1)} = True, X_{\{0.33,0.66\}}^{(2)} = False\} \implies \{\text{Event=True}\}$	0.2315789	0.9777778	1.955556
$\{X_{\{0,0.33\}}^{(1)} = True, X_{\{0.33,0.66\}}^{(3)} = False\} \implies \{\text{Event}=\text{True}\}$	0.2236842	0.9770115	1.954023
$\{X_{\{0,0.33\}}^{(1)} = True, X_{\{0,0.33\}}^{(4)} = False\} \implies \{\text{Event}=\text{True}\}$	0.2210526	0.9767442	1.953488
${X_{\{0,0.33\}}^{(5)} = True, X_{\{0.33,0.66\}}^{(2)} = False} \implies {\text{Event=True}}$	0.3118421	0.9753086	1.950617
$\{X_{\{0,0.33\}}^{(1)} = True, X_{\{0.33,0.66\}}^{(6)} = False\} \implies \{\text{Event=True}\}$	0.2052632	0.9750000	1.950000
$\{X^{(5)}_{\{0,0.33\}} = True, X^{(2)}_{\{0,66,1\}} = False\} \implies \{\text{Event=True}\}$	0.2552632	0.9748744	1.949749
$\{X_{\{0,0.33\}}^{(1)} = True, X_{\{0.33,0.66\}}^{(7)} = False\} \implies \{\text{Event=True}\}$	0.2013158	0.9745223	1.949045
$\{X_{\{0,0.33\}}^{(5)} = True, X_{\{0,0.33\}}^{(3)} = False\} \implies \{\text{Event}=\text{True}\}$	0.2355263	0.9728261	1.945652
$\{X_{\{0,0.33\}}^{(1)} = True, X_{\{0,0.33\}}^{(2)} = False\} \implies \{\text{Event}=\text{True}\}$	0.2315789	0.9723757	1.944751
$\{X_{\{0,0,33\}}^{(1)} = True, X_{\{0,33,0.66\}}^{(8)} = False\} \implies \{\text{Event=True}\}$	0.2315789	0.9723757	1.944751

Table 3: This table displays the rules found by the algorithm sorted by confidence and lift. The support is also shown here.

approach CAPP2, add variables based on mean thresholds, improve our predictive model by aggregating the results over multiple analyses with different Δ and δ and optimize the event detection system.

ACKNOWLEDGMENTS

This work was supported by TotalEnergies and the research laboratory SINCLAIR. We also thank Tobias Kurth, Professor of Public Health and Epidemiology and the Director of the Institute of Public Health at the Charité Universitätsmedizin in Berlin, and Lluvia Ochoa, Researcher in Data Science and Artificial Intelligence at TotalEnergies who provided their insights and expertise.

REFERENCES

- H Alkharusi. 2012. Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education* 4(2) (2012), 202–210.
- [2] Jorge A.M. Azevedo P.J. 2007. Comparing Rule Measures for Predictive Association Rules. ECML (2007).
- [3] Leo Breiman. 2001. Random Forests. Machine Learning 45 (2001), 5–32.
- [4] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. 2017. Classification And Regression Trees. 1–358 pages. https://doi.org/10.1201/9781315139470
- [5] Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. 2021. SIRUS: Stable and Interpretable RUle Set for classification. *Electronic Journal of Statistics* 15 (01 2021), 427–505. https://doi.org/10.1214/20-EJS1792
- [6] Yonghong Chen, Govindan Rangarajan, Jianfeng Feng, and Mingzhou Ding. 2004. Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A* 324, 1 (2004), 26 – 35. https://doi.org/10.1016/j.physleta.2004.02.032
- [7] K. Dembczynski, W. Kotlowski, and R. Slowinski. 2010. ENDER: a statistical framework for boosting decision rules. *Data Mining and Knowledge Discovery* 21 (2010), 52–90.
- [8] Doris Entner and Patrik Hoyer. 2010. On Causal Discovery from Time Series Data using FCI. Proceedings of the 5th European Workshop on Probabilistic Graphical Models, PGM 2010 (09 2010).
- [9] Gardner IA et al. Estberg L. 1998. A case-crossover study of intensive racing and training schedules and risk of catastrophic musculoskeletal injury and lay-up in California thoroughbred racehorses. *Prev. Vet. Med* 33 (1998), 159–170.
- [10] Maclure M et al. 1993. Triggering of acute myocardial infarction by physical exertion: protection against triggering by regular exertion. N. Engl. J. Med 329 (1993), 1677–1683.
- [11] Maclure M et al. 1995. Triggering of acute myocardial infarction by episodes of anger. Circulation 92 (1995), 1720–1725.
- [12] A. Eye, Wolfgang Wiedermann, and I. Koller. 2015. Granger Causality: Linear Regression and Logit Models.
- [13] Eibe Frank and Ian Witten. 1998. Generating Accurate Rule Sets Without Global Optimization. Machine Learning: Proceedings of the Fifteenth International Conference (06 1998).
- [14] C. W. J. Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37, 3 (1969), 424–438. http://www. jstor.org/stable/1912791

- [15] T. Hastie, R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. In Springer Series in Statistics.
- [16] Liang H. Hu M. 2014. A copula approach to assessing Granger causality. NeuroImage 100 (2014), 125–134.
- [17] T. Imielinski, Arun Swami, and Rajat Agrawal. 1993. Mining Association rules between sets of items in large databases. ACM SIGMOD (01 1993), 207–216.
- [18] J. R KISTER, H. Z; HAAS. 1990. Predict entrainment flooding on sieve and valve trays. *Chemical engineering progress* (1990).
- [19] O. Ludwig, U. Nunes, R. Araújo, L. Schnitman, and H. Lepikson. 2009. Applications of information theory, genetic algorithms, and neural models to predict oil flow. *Communications in Nonlinear Science and Numerical Simulation* 14 (2009), 2870– 2885.
- [20] M. Maclure. 1991. The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events. *American journal of epidemiology* 133 (1991), 144–153. https://doi.org/10.1093/oxfordjournals.aje.a115853
- [21] M. Maclure and M. Mittleman. 1997. Cautions about car telephones and collisions. The New England journal of medicine 336 7 (1997), 501–2.
- [22] Mittleman MA Maclure M. 1997. Cautions about car telephones and collisions. N. Engl. J. Med. 336 (1997), 501–502.
- [23] Stramaglia S. Marinazzo D., Pellicoro M. 2008. Kernel method for nonlinear Granger causality. Phys. Rev. Lett. 100 (2008), 144103.
- [24] R. Michalski. 1969. On the quasi-minimal solution of the general covering problem. (01 1969).
- [25] Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal Discovery with Attention-Based Convolutional Neural Networks. *Machine Learning and Knowl-edge Extraction* 1, 1 (2019), 312–340. https://doi.org/10.3390/make1010019
- [26] J. Pearl. 2000. Causality: Models, Reasoning and Inference. Cambridge University Press.
- [27] R. Srikant R. Agrawal. 1994. Fast Algorithms for Mining Association Rules, In VLDB Conference. Mathematical Modelling and Computational Experiments.
- [28] Dilini Rajapaksha, Christoph Bergmeir, and Wray Buntine. 2019. LoRMIKA: Local Rule-based Model Interpretability with k-optimal Associations. arXiv:1908.03840 [cs.LG]
- [29] Sebastian Raschka. 2014. Mlxtend Apriori. http://rasbt.github.io/mlxtend/user_ guide/frequent_patterns/apriori/.
- [30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]
- [31] Jakob Runge, Dino Sejdinovic, and Seth Flaxman. 2017. Detecting causal associations in large nonlinear time series datasets. *Science Advances* 5 (02 2017). https://doi.org/10.1126/sciadv.aau4996
- [32] S. Salzberg. 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning 16 (1994), 235–240.
- [33] Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. Causation, Prediction, and Search, 2nd Edition. MIT Press Books, Vol. 1. The MIT Press. https://ideas. repec.org/b/mtp/titles/0262194406.html
- [34] Giulia Toti, R. Vilalta, P. Lindner, B. Lefer, Charles Macias, and D. Price. 2016. Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. *Artificial intelligence in medicine* 74 (2016), 44–52.
- [35] Sholom Weiss and Nitin Indurkhya. 2000. Lightweight Rule Induction. (05 2000).