Journal of Biomedical Informatics 108 (2020) 103512

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

An analysis of malaria in the Brazilian Legal Amazon using divergent association rules

Lais Baroni[a], Rebecca Salles[a], Samella Salles[c], Gustavo Guedes[a], Fabio Porto[d], Eduardo Bezerra[a], Christovam Barcellos[b], Marcel Pedroso[b], Eduardo Ogasawara[a,*]

[a] *CEFET/RJ, Brazil*
[b] *FIOCRUZ, Brazil*
[c] *IFRJ, Brazil*
[d] *LNCC, Brazil*

ARTICLE INFO

Keywords:
Pattern mining
Malaria
Divergent association rules

ABSTRACT

In data analysis, the mining of frequent patterns plays an important role in the discovery of associations and correlations between data. During this process, it is common to produce thousands of association rules (ARs), making the study of each one arduous. This problem weakens the process of finding useful information. There is a scientific effort to develop approaches capable of filtering interesting patterns, balancing the number of ARs produced with the goal of not being trivial and known by specialists. However, even when such approaches are adopted, the number of produced ARs can still be high. This work contributes by presenting Divergent Association Rules Approach (DARA), a novel approach for obtaining ARs that presents themselves in divergence with the data distribution. DARA is applied right after traditional approaches to filtering interesting patterns. To validate our approach, we studied the dataset related to the occurrence of malaria in the Brazilian Legal Amazon. The discovered patterns highlight that ARs brought relevant insights from the data. This article contributes both in the medical and computer science fields since this novel computational approach enabled new findings regarding malaria in Brazil.

1. Introduction

The mining of frequent patterns plays an important role in the discovery of associations and correlations between data [1]. Patterns that are frequent in a dataset can be expressed by association rules (ARs). ARs highlight frequent itemsets in the antecedent leading to those in the consequent [2].

During the mining of frequent patterns, it is common to produce thousands of ARs, making the study of each one arduous. This problem weakens the process of discovery of useful information. There is a scientific effort to develop approaches capable of filtering interesting patterns, balancing the number of ARs produced with the goal of not being trivial and known by specialists. Among the approaches for filtering interesting patterns, some use measures of interest [3,4], others list them based on properties [5–7], and others use subjective analysis [8–10].

The difficulty of filtering interesting patterns increases with the number of generated patterns. It is strongly influenced by the potentially high cardinality of attributes available in the dataset and by choice of the minimum support to be considered during mining [11]. Thus, even though the literature presents a diversity of approaches and the possibility of combining several of them, they are not always able to achieve the previously stated goal [12].

For example, consider the evolution of malaria in the Amazon region. Malaria is an infectious disease caused by protozoan parasites of the genus *Plasmodium* and transmitted through the bite of the mosquito of the genus *Anopheles* [13]. Tropical and subtropical countries comprise the endemic area of this disease. They have rainy seasons that provide high availability of clean standing water, where the vector mosquitoes can lay their eggs and proliferate [14]. In this context, Brazil is the second country in the Americas with the highest number of malaria cases [15]. According to the World Health Organization, analyzing associations of variables related to positive outcomes for malaria is essential for its control and support in the decision-making process [15].

The Malaria Epidemiological Surveillance Information System (SivepMalaria) is one of the main systems for monitoring malaria in Brazil. Through this system, all suspected or confirmed cases are

[*] Corresponding author.
E-mail address: eogasawara@ieee.org (E. Ogasawara).

https://doi.org/10.1016/j.jbi.2020.103512
Received 18 February 2020; Received in revised form 9 July 2020; Accepted 10 July 2020
Available online 21 July 2020

notified and recorded. However, the occurrence of the disease is not homogeneous, varying from location to location according to some characteristics such as natural factors, geographic factors, and social conditions [16,17].

By using classic approaches to discover frequent patterns in datasets (such as SivepMalaria), thousands of ARs can be produced; however, several are redundant or not interesting. Even using traditional approaches to filter interesting ARs, the number of ARs can still be high. This work proposes the Divergent Association Rules Approach (DARA), which is a novel approach to filtering interesting ARs from the divergence between the obtained ARs and the data distribution. DARA focuses on the items present in the ARs at a lower or higher frequency than the frequency expected considering the data distribution. They lead to Divergent Association Rules (DARs). DARs assumes the hypothesis of attributes independence. Thus, such divergence in the frequency is not expected to occur. When it occurs, DARs might be interesting.

This paper has two main contributions. In the computer science field, it presents DARA, a novel approach to obtaining DARs, which are ARs that are considered divergent concerning data distribution. The approach both sheds light on interesting patterns and drives a systematic way of discovering them. Furthermore, it serves as a complement to other approaches to filtering interesting ARs.

The second contribution of this paper is related to the medical field, as DARA was evaluated on data from SivepMalaria. The approach was able to highlight ARs that bring relevant insights from the input data. New patterns regarding Brazilian Legal Amazon are revealed.

In addition to this introduction, this paper is organized into five more sections. Section 2 presents background on frequent pattern mining. Section 3 describes related works. Section 4 presents DARA. Section 5 presents the experimental evaluation. Section 6 presents some concluding considerations.

2. Background

Techniques for frequent pattern mining aim at supporting the process of extracting knowledge to discover relevant associations between items. It starts from the observation of frequent items, *i.e.*, items that occur at least as frequently as a predetermined minimum support count [18]. Formally, an item X_i corresponds to the association between an attribute X with a value x_i. An itemset A_p corresponds to a set of items, such that $A_p = \{X_{p_1}, Y_{p_2}, \cdots, Z_{p_n}\}$, for $n = |A_p|$.

Given two frequent itemsets A_p and B_q in a dataset, they can be related through an association rule of the form $A_p \Rightarrow B_q$. The ARs work in a way that presents the frequent itemset in the antecedent (LHS) or consequent (RHS) [2]. The itemset in the antecedent is the necessary condition to reach the itemset of the consequent. Considering the malaria dataset, for example, the rule "*gender* = *male*⇒*plasmodium* = *falciparum*". It is represented as: if the patient's gender is male (antecedent), it leads to *plasmodium* being *falciparum* (consequent)[1].

ARs are generated in two steps during frequent pattern mining. The first step computes all the frequent itemsets in the dataset, and the second one generates the ARs from the previously found frequent itemsets. Several algorithms have been developed (and improved) for this purpose [19], among them we can highlight Apriori [20], FP-growth [21], and ECLAT [22]. However, constraints are necessary to reduce the number of possible ARs. These constraints can be given as input to the ARs creation algorithm and are commonly the minimum and maximum size of the ARs and the minimum thresholds for support and confidence. These constraints limit the set of ARs generated, as these are conditions to be met.

The support of an item X_i is the frequency of X_i within the dataset [23]. Formally, given a dataset D, the support (*sup*) for an item X_i is the probability of occurrence of the event $X = x_i$ in D (*i.e.*, $sup(X_i) = \Pr(X = x_i)$). For example, if the value 50% is defined as the threshold for support, then only items that appear in at least half of the transactions [21] are considered as frequent. Analyzing an association rule of type $A_p \Rightarrow B_q$, we have that $sup(A = a_p \Rightarrow B = b_q)$ is the probability that both itemsets will occur together, *i.e.*, $sup(A_p \Rightarrow B_q) = sup(A_p \cup B_q) = \Pr(A = a_p \cap B = b_q)$[2]. If an item appears less frequently than the support determined in the algorithm for ARs generation, this item does not appear in any ARs. On the other hand, setting low support thresholds can lead to the generation of more ARs than would be useful in practice [25]. In short, low supports result in a higher number of ARs. On the other hand, high support values can remove less frequent items from the analysis, which could be of interest.

Besides support, another necessary input constraint for generating ARs is confidence. Confidence does not care about the frequency of the item in the dataset. Instead, it is related to the chance that the rule is valid. It is a measure of the probability of occurrences of the consequent, given that the antecedent happens [23]. In an association rule $A_p \Rightarrow B_q$, its confidence is given by $conf(A_p \Rightarrow B_q) = \Pr(A_p|B_q)$, representing the conditional probability that B_q given A_p. In the rule previously given as an example, if it has the confidence of 80%, it means that of all transactions in which item *plasmodium* = *falciparum* appears, 80% also have item *gender* = *male*. Therefore, while support expresses the frequency of rule itemsets, confidence is a measure of the strength of the association between these itemsets.

2.1. Interesting ARs

One of the difficulties in extracting knowledge from ARs is that a vast number of ARs can be created even from a small set of transactions. Another obstacle is the fact that many of the strong ARs (ARs that satisfy both minimum support and confidence thresholds) found are trivial and uninteresting [26]. Therefore, extracting interesting ARs is a challenging task [27].

Subjective approaches used to reduce frequent patterns, require a more in-depth knowledge of the database that is mined and, ideally, the participation of an expert in the domain [28,9,10,29]. Thus, one can determine, within the scope of the research, those ARs or values that are not interesting to be analyzed. Often values such as 'other' or 'ignored' are then removed from the analysis.

An association rule can also be measured by the correlation between itemsets in its antecedent and consequent. They are called interesting measures [1,30]. Among the abundant amount of measures disseminated in the literature, there are *lift*, *Kulczynski*, and *imbalance ratio* [1].

The *lift* for a rule $A_p \Rightarrow B_q$ is defined by Eq. (1).

$$lift(A_p \Rightarrow B_q) = \frac{conf(A_p \Rightarrow B_q)}{sup(B_q)} \tag{1}$$

This measure evaluates the degree to which the occurrence of one itemset promotes the occurrence of the other [31]. A value of *lift* equal to one indicates the independence between itemsets. A value higher than one indicates an association between them, or in other words, that they are complementary (or are positively related) [32]. The *lift* below one means that the presence of one itemset inhibits the presence of the other (i.e., they are negatively related). In this case, the itemsets are considered complementary.

[1] In this example, the itemsets in the antecedent and consequent are composed of only one item to easy understanding.

[2] In pattern mining, $A_p \cup B_q$ correspond to the union of itemsets [1,11], *i.e.*, the union of constraints over dataset D. In statistics, such constraint corresponds to the intersection of their respective events over D [24], *i.e.*, $\Pr(A = a_p \cap B = b_q)$.

The *kulc* for a rule is defined by the Eq. (2) [1].

$$kulc(A_p \Rightarrow B_q) = \frac{\Pr(A_p|B_q) + \Pr(B_q|A_p)}{2} \tag{2}$$

This degree of correlation is represented by a real number between 0 and 1. If *kulc* is close to either zero or one, then the rule is considered interesting, being negative or positive associated, respectively. If the resulting number is close to 0.5, the *kulc* index is considered neutral, *i.e.*, the rule may or may not be interesting.

The *imbalance ratio* (*IR*) for a rule $A_p \Rightarrow B_q$ is expressed by the Eq. (3).

$$IR(A_p \Rightarrow B_q) = \frac{|sup(A_p) - sup(B_q)|}{sup(A_p) + sup(B_q) - sup(A_p \Rightarrow B_q)} \tag{3}$$

IR measures the degree of asymmetry between two events that contain antecedent itemset A_p and the consequent itemset B_q [33]. The numerator is the absolute value of the difference between the supports of the A_p and B_q itemsets and the denominator is the percentage (support) number of transactions that contain A_p or B_q. However, it does not contain both together.

Both *kulc* and *IR* have a null-invariance property [1]. A typical application is to make use of these measures together. It consists of the idea of first filtering the interesting ARs by using the *kulc* measure and then the *IR* measure to evaluate the ARs that presented *kulc* close to 0.5 (neutral). Since neutral *kulc* is not very informative, an *IR* (balanced close to zero) indicates an uninteresting rule. If the *IR* shows values close to one, then the rule can be considered interesting [1].

Besides measures of interest, another resource for studying interesting patterns is based on the consideration of redundancies in the set of ARs. The logic involved in the decision process is based on the idea that a rule is redundant if there are more general ARs with the same or higher confidence. A more general rule means a rule with the same consequent (RHS) and a smaller number of items in the antecedent (the items present being equal to those of the less general ARs). Thus, if adding an item to the left side of the rule leads to a decrease in the confidence, or it remains the same, then that rule is redundant. It means that the consideration of that new item is equivalent to a negative or null improvement in the rule [6]. Formally, we consider that a rule $A_p \Rightarrow B_q$ is redundant if $\exists A'_r \subset A_p | conf(A'_r \Rightarrow B_q) \geqslant conf(A_p \Rightarrow B_q)$.

2.2. Mining process for producing frequent patterns

From the concepts introduced so far, it is possible to present the general Data Mining process for frequent patterns described in Algorithm 1. The algorithm receives as input a transaction dataset *D*, a support threshold *sup* and confidence *conf*, a set of constraints *cons*, and a set of approaches to obtain interesting ARs *inter*. The output of the algorithm is the set *R* with all the ARs filtered by constraints and considered interesting.

Algorithm 1. Mining of Frequent Patterns

```
1: input: dataset $D$; support $sup$; confidence $conf$; constraints $cons$; interesting measures $inter$.
2: output: Set of ARs $ICR$.
3: function $pattern\_miningD, sup, conf, cons, inter$
4:    $P \leftarrow apriori(D, sup)$
5:    $R \leftarrow gen\_rules(P, conf)$
6:    $R \leftarrow apply\_constraints(R, cons)$
7:    $R \leftarrow apply\_interestingness(R, inter)$
8:    return $R$
```

In the line 4, the *apriori* algorithm takes the dataset (*D*) and minimal support (*sup*) as a parameter. It produces *P*, the set of frequent patterns. Then, the *P* set together with the minimum confidence parameter (*conf*), is used by the *gen_rules* algorithm and produces the set of ARs *R* (line 5).

In the line 6, some constraints are applied to these ARs. The constraints can be values for LHS and RHS, minimum and maximum sizes for ARs. Such constraints are meant to reduce the set of ARs *R* from the previous ones.

Finally, in the line 7, interesting rule selection measures are applied to *R*, such as, for example, the indication of suitable values for *lift*, *kulc* and non-redundant ARs. In the context of this paper, DARA fits as a new approach to obtaining interesting ARs that can be combined with the other existing ones. This step produces a final rule set *R* returned by Algorithm 1 (line 8).

3. Related work

A well-known problem in pattern mining is that frequent pattern discovery methods generate hundreds and often thousands of ARs. The number of produced ARs makes the analysis unfeasible, weakening the process of discovering useful information. An important task, then, is to determine the most useful patterns among them, that is, those that are not trivial or already known. In the literature, several papers address methods to determine sets of interesting patterns in a dataset. The methodology of frequent patterns and ARs is already a way of listing interesting patterns using support and confidence measures.

In addition to support and confidence, several other measures of interest play an important role in this context. They can discover dependencies and correlations between variables in a dataset and enable the filtering of patterns according to their values. Some works create or derive new measures of interest [3,4]. Other works study measures to describe patterns utilities and applications [12,34]. Other approaches were developed to assist in the task of selecting the most relevant patterns from all those generated. Some examples are the use of closed patterns [5], maximum patterns [35], redundant patterns [6], and emerging patterns [7]. All of these are concepts widely used and disseminated in the literature.

Some works have developed approaches to filter unexpected ARs [8,28,9,10,36,37,29]. They consider the opinions of experts in the field that guide the process of identifying expected patterns. From the expert's definition, unexpected ARs are discovered according to criteria of interest.

Other authors, however, propose different approaches to the treatment of available patterns. Gan et al. [38] and Fournier et al. [39–41] propose algorithms that associate the utility of the patterns and the correlation between the items in the pattern. Soulet et al. [42] aimed at useful patterns considering user preference. They work with the idea of *skyline* queries to extract interesting *skyline* patterns.

In the context of relevant patterns, Yan et al. [43] developed a methodology to compress patterns that are generated from data streams. It is suitable to continuously filter representative patterns in sliding window event streams using the minimum description length. Pellegrina et al. [44] focus on the process of finding statistically significant patterns. Significance is usually given from a statistical test that quantifies the probability that the association observed in real data appears only by chance.

To the best of the authors' knowledge, no other work in the literature considers the question of comparing the frequency of occurrences of the items in the set of discovered ARs and the dataset, as proposed by DARA.

4. Methods

This paper presents DARA, a novel approach to discover DARs. They are named DARs when there exists at least one item in the antecedent such that their presence in the dataset is not in agreement with their presence in ARs. It considers the hypothesis of attributes independence.

4.1. Formalization

Lemma 1. *Suppose two items X_i and X_j associated with the same attribute X, where X_i has higher support than X_j, i.e.,* $\Pr(X = x_i) > \Pr(X = x_j)$. *Assuming the hypothesis of attributes independence, a minimum support σ_{min}, a minimum confidence δ_{min} and a consequent Y_k, during the generation of ARs that leads to Y_k whichever X_i and X_j are in the antecedent, the probability of having ARs containing X_i is higher than X_j.*

To prove Lemma 1, let us analyze ARs from the support perspective. Suppose there is a rule $A_p \Rightarrow Y_k$, where Y_k is a consequent itemset and A_p correspond to itemset such that $X_i, X_j \nsubseteq A_p$. This rule has support equal to $sup(A = a_p \Rightarrow Y = y_k) = Pr(A = a_p \cap Y = y_k) = \sigma$, with σ higher than the defined minimum support (σ_{min}). Also, consider that the supports for X_i and X_j are higher than the defined minimum support (σ_{min}). If the item X_i is added to the set A_p (for simplicity, denoted A_pX_i) for the formation of the rule $A_p X_i \Rightarrow Y_k$, $sup(A_pX_i \Rightarrow Y_k) = \sigma_i = Pr(A = a_p \cap X = x_i \cap Y = y_k)$. Assuming the hypothesis of attributes independence, we have that: $Pr(A = a_p \cap X = x_i \cap Y = y_k) = Pr(A = a_p \cap Y = y_k) \times Pr(X = x_i)$ Similarly, adding X_j to the set A_p for the formation of the rule $A_pX_j \Rightarrow Y_k$, we have $sup(A_pX_j \Rightarrow Y_k) = \sigma_j = Pr(A = a_p \cap Y = y_k) \times Pr(X = x_j)$.

For an ARs to be formed with the item X_i, then the support must be higher than the minimum support. Therefore, it becomes necessary that $\sigma \times \sigma_i > \sigma_{min} \therefore \Pr(X = x_i) > \frac{\sigma_{min}}{\sigma}$. Likewise, considering the item X_j, we have that $\Pr(X = x_j) > \frac{\sigma_{min}}{\sigma}$. Since $\Pr(X = x_i) > \Pr(X = x_j)$, $\frac{\sigma_{min}}{\sigma}$ can admit a higher value in the rule associated with X_i when compared to X_j.

Therefore, more ARs can be associated with X_i than with X_j. Remember that σ_{min} is fixed for all ARs and σ depends on the items present in $A_p \Rightarrow Y_k$. Therefore, it can be expected that the number of ARs formed with an item X_i is proportional to the support of X_i in the dataset. Furthermore, considering the evidence given, it is possible to derive the Corollary 1 from Lemma 1.

Corollary 1. *Consider the hypothesis of attributes independence, a minimum support σ_{min}, and a minimum confidence δ_{min} to generate the ARs $A_p \cup X_i \Rightarrow Y_k$ and $A_p \cup X_j \Rightarrow Y_k$. If the probability of the occurrence of two items X_i and X_j referring to an attribute X are equal, then these two items are contained in the same number of ARs.*

Let us now analyze Lemma 1 from the perspective of confidence and assuming the hypothesis of attributes independence. We have that $conf\,(A_p \Rightarrow Y_k) = \Pr(Y = y_k | A = a_p) = \frac{\Pr(A = a_p \cap Y = y_k)}{\Pr(A = a_p)} =$. That is,

$$\frac{\Pr(A = a_p) \times \Pr(Y = y_k)}{\Pr(A = a_p)} = \Pr(Y = y_k)$$

concerning confidence, as long as the minimum confidence δ_{min} remains fixed, there is no influence on the number of ARs generated when adding a new item on the left side of the rule (A_p). Thus, it is possible to derive the Corollary 2 from Lemma 1.

Corollary 2. *Assume the hypothesis of attributes independence, a minimum support σ_{min}, and a minimum confidence δ_{min} to generate ARs like $A_p \Rightarrow Y_k$. If the consequent Y_k is fixed, then the confidence of the rule is not affected by the change in the itemset A_p in the antecedent.*

4.2. DARA: Divergent association rules approach

The divergent items for an attribute X are those found by comparing the frequencies of the items X_i and X_j in X concerning the number of ARs generated involving them (in the antecedent) targeting a consequent item Y_k. When Lemma 1 is not observed, it means that a more frequent item X_i generates lower ARs than a less frequent item X_j or vice versa. Item X_i presents divergence, and ARs containing X_i are considered DARs.

Algorithm 2 presents DARA. The algorithm takes the data set D as input, the itemset Y_k that is the right side of the ARs and the set of ARs R that lead to Y_k. For each attribute X in D (line 5), a paired comparison is made between its items (X_i and X_j) (line 6) according to Lemma 1 and Corollary 1. The support of each pair of items X_i and X_j is calculated for the attribute X in D (lines 7 and 8). It corresponds to $\Pr(X = x_i)$ and $\Pr(X = x_j)$. The ARs in which X_i and X_j are on the left side of the rule are also observed (lines 9 and 10). If the number of ARs containing X_i is less than the number of ARs containing X_j and $\Pr(X = x_i)$ is higher than $\Pr(X = x_j)$, there is a negative divergence between X_i and X_j (lines 11 and 12). Conversely, if the number of ARs containing X_i is higher than the number of ARs containing X_j and $\Pr(X = x_i)$ is less than $\Pr(X = x_j)$, there is a positive divergence between X_i and X_j (lines 14 and 15). Positive and negative divergences are stored in a counter L_{X_i} (line 17). In the end, values of counter L_{X_i} that present the highest absolute values of divergences are driven for analysis. As stated in Corollary 2, the confidence of the ARs is not taken into account in Algorithm 2.

Algorithm 2. Detection of divergent items

```
1: input:$D$: dataset, $R$: ARs that lead to a consequent $Y_k$
2: output: The set of positive and negative divergences
3: function $divergenceD, R$
4:   $L \leftarrow \emptyset$
5:   for $X \in attributes(D)$ do
6:     for $X_i, X_j \in attribute(X) | X_i \neq X_j$ do
7:       $p_{X_i} \leftarrow sup(X_i)$
8:       $p_{X_j} \leftarrow sup(X_j)$
9:       $r_{X_i} \leftarrow rules(R, X_i)$
10:        $r_{X_j} \leftarrow rules(R, X_j)$
11:        if $(|r_{X_i}| < |r_{X_j}| \wedge p_{X_i} > p_{X_j})$ then
12:          $L_{X_i} \leftarrow L_{X_i} - 1$
13:        end if
14:        if $(|r_{X_i}| > |r_{X_j}| \wedge p_{X_i} < p_{X_j})$ then
15:          $L_{X_i} \leftarrow L_{X_i} + 1$
16:        end if
17:      return {L};
```

Table 1 exemplifies a hypothetical case to easy understanding. Assuming a ‘color’ attribute with three possible values: ‘white’, ‘black’, ‘gray’. The number of times each color appears in the dataset is shown in the ‘Freq.’ column. The number of ARs in which that same color appears is shown in the ‘ARs’ column.

In this example, the counter for the ‘white’ and ‘black’ values will be + 1, and the counter for the ‘gray’ value will be − 2 since it has fewer ARs compared to the other two values. Thus, the ‘gray’ value is an example of a divergent value. It is noticed that the values assigned to the counters of the values ‘white’ and ‘black’ are a reflection (compensation) of the divergence of ‘gray’.

5. Experimental evaluation

The experimental evaluation is divided into dataset setup (Section 5.1), experimental setup (Section 5.2), quantitative analysis (Section 5.3), qualitative analysis (Section 5.4), and discussion (Section 5.5).

5.1. Dataset setup

Given the need to obtain useful knowledge about malaria, an integrated dataset of malaria notifications in the Legal Amazon was

Table 1
Hypothetical example for determining divergent values.

Value	Freq.	ARs
white	100	10
black	200	20
gray	300	5

used[3]. It is a dataset with all medical records of SivepMalaria about patients who were tested for malaria in the Brazilian Legal Amazon.

The information system consists of modules that record notification data, examination data, and information about patients who have undergone the examination for malaria in the states of the Legal Amazon [45]. For this study, data from seven years (2009 to 2015) were considered, comprising 15,764,287 records. About 12% of these records corresponds to positive cases of malaria, and only in these cases, information about the patient is filled out.

Some modifications were needed to prepare the integrated dataset of SivepMalaria for the mining of frequent patterns. Regarding the 40 attributes available for analysis, some of them were not relevant for the analysis. They were removed, such as **qty.parasites**, **scheme**, **pregnancy**, **crosses**, and **cvl.case**.

Regarding the attributes related to space, the health region was the spatial unit used for analysis. These regions are part of the systemic organization of public health, aiming at political-administrative decentralization and comprehensive care. They constitute a homogeneous region of observation in the context of health. Spatial attributes included in the analysis were **home.hr**, **migration**, and **autochthonous.case**. The **home.hr** attribute was renamed **address** when it was enriched with the cases where the residence of the patient is outside the Brazilian states of the Legal Amazon. The values from the attributes **home.state** and **home.country** were used for that. The states outside the Legal Amazon were mapped into two categories: "border state" and "other states". The "border state" stands for those states that border any of the nine states of the Legal Amazon. The "other states" represents the other states of Brazil. Similarly, the country attribute was mapped to "border country" for the countries that border with the states of the Legal Amazon. All other spatial attributes were removed from the analysis.

Regarding the attributes related to time (**exam.interval**, **treatment.interval**, and **notification.interval**), only the attributes of month and year of notification were kept, *i.e.*, the attributes **notification.month** and **notification.year**. They enable (based on the ARs) to observe the simultaneous effects of seasonality and abnormal years. All other dates were removed from the analysis[4].

5.2. Experimental setup

The ARs were generated using the Apriori algorithm, available through the package Arules [46] of the software R [47]. The research objective is to understand the factors associated with the occurrence (or non-occurrence) of malaria. Thus, the consequent of the ARs (RHS) was set up to the examination result attribute. Possible values include *Plasmodium falciparum* (Falciparum), Non-Falciparum, *Plasmodium vivax* (Vivax), *Plasmodium malariae* (Malariae), *Plasmodium ovale* (Ovale), and Negative (when malaria was not confirmed).

As the frequency of each examination result is very different, the dataset was partitioned according to it. The support of ARs in each resulting part (for short, part) was defined by the maximum curvature method [48], computed according to the frequency of five attributes: month (**notification.month**), year (**notification.year**), residence (**address**), gender (**gender**), and age (**age**) of the patient.

Table 2 presents the support values used for the generation of ARs in each part. As can be seen, it was not possible to compute the support for the ovale malaria type. This is because only nine cases of this ovale type occur in the entire dataset, which is not enough for pattern mining analysis. The absolute and relative supports are, respectively, the support calculated by the maximum curvature method, and this value divided by the number of rows of the dataset.

Table 2
Support definition for each part.

RHS	Absolute Support	Relative Support
Negative	161	1.02×10^{-5}
Vivax	149	9.4×10^{-6}
Falciparum	33	2.1×10^{-6}
Non-Falciparum	10	6.3×10^{-7}
Malariae	2	1.2×10^{-7}
Ovale	-	-

Table 3 summarizes the values of parameters common to all parts. The place of residence attribute was enforced in the antecedent of the ARs (LHS). The minimum confidence was defined as 80%. The confidence adopted was high to ensure that the itemset of the antecedent (LHS) are decisive to promote the result presented on the right side of the rule (RHS). Finally, the processing time has been left unlimited, causing the processing to end only when all possible ARs are built.

5.3. Quantitative analysis

Once patterns are produced, the approaches for filtering interesting ARs were applied to provide a "cleaner" analysis. The approaches used in this work for filtering purposes were based on (i) irrelevant values, (ii) measures of interest, and (iii) non-redundancy. The filtering of irrelevant information was based on expert opinion. The disregarded values were: "occupation = ignored", "occupation = other", "scholarship = not applicable", "autochthonous.case = yes", "treatment.interval = on the same day", "exam.interval = on the same day", "previous.treatment = no", and "migration = no".

Then, the filtering based on measures of interest included *lift*, *kulc*, and *imbalance ratio*. ARs deemed uninteresting according to these measures were removed (as explained in Section 2.1). Finally, all ARs considered redundant have also been removed from the ARs sets.

Table 4 summarizes the number of ARs generated using the Apriori algorithm for each subset of SivepMalaria. The result of filtering interesting ARs using three different traditional approaches is presented (*nr*: based on irrelevant values, *i*: based on measures of interest, *r*: based on non-redundancy). The result considering the combination of the three approaches (filtering interesting ARs) is also presented. Finally, the number of ARs filtered by DARA is computed after traditional approaches.

As it can be observed, when the number of ARs found by the combined approaches is still high, DARA can significantly reduce the number of ARs to study. Also, the malariae malaria type is not presented in Table 4 since no AR was generated for this subset. Appendix B shows the DARA results for each of these sets, with an indication of the divergent values.

5.4. Qualitative analysis

In this subsection, the analyses made with DARA in the malaria domain are presented. It includes analysis of attributes of notification year, type of hemoparasites, occupation, regions of health, and race.

It is worth mentioning that these results were found based on the specific study of the divergent items pointed out by the DARA approach. DARA works on a subset of previously computed association rules. The same results could be found directly studying the entire ARs subset, without using the DARA. However, it might be harder to find them. Also, since DARA is associated with divergent items, they have grouped accordingly, which might ease their discovery.

5.4.1. Divergence in the notification year attribute

The first interesting observation was the perception that the years 2009 and 2010, which are the ones that most appear in the

[3] https://www.synapse.org/#!Synapse:syn21609131.

[4] Table 8 (see Appendix A) presents the 19 attributes of the preprocessed dataset used during experimental evaluation.

Table 3
Common parameters for the generation of ARs in all parts.

Parameter	Assigned Value
LHS	**address**
minimum size	3
maximum size	4
confidence	80%
maximum time	∞

Table 4
The number of ARs generated, the number of ARs filtered by traditional approaches (*nr*: based on irrelevant values, *i*: based on measures of interest, *r*: based on non-redundancy), and the number of ARs filtered by DARA (after traditional approaches).

	Negative	Vivax	Falciparum	Non-Falciparum
Generated	1, 452	64, 104	44	407
Filtered by *nr*	316	23, 461	30	283
Filtered by *i*	608	64, 104	44	407
Filtered by *r*	849	28, 567	34	329
Filtered by *nr*, *i* and *r*	118	10, 838	24	224
Filtered by *nr*, *i*, *r* and DARA	89	2, 761	24	93

SivepMalaria dataset, appear less frequently in the ARs. Hence, we looked for a possible reason for this.

This result was novel, as it was not previously documented in the available dataset [49]. Thus, the preprocessed dataset was studied again, focusing on the notification year attribute. It was noted that some attributes in the records of these two years were rarely filled out, resulting in many empty fields. It disabled the generation of ARs for those years. It represents an interesting observation of data completion and completeness that was not seen in the initial exploratory analysis.

The missing value attributes are those that presented information about researched hemoparasites, type of examination, previous treatment, and race of the patient. These four attributes started to be registered only in 2011, when the SivepMalaria form was changed [45].

DARA helped us to check the consistency of the dataset. Table 5 shows the relative supports indicating how much of the records are related to the ARs generated for the patients infected with *P. vivax*. The support is high for every year, varying around 70%, including the years 2009 and 2010, where the divergence is found. Therefore, if ARs were only analyzed according to their support, nothing could be observed regarding the peculiarity of the missing attributes in the first two years considered in this study.

This evaluation enabled us to conclude that fewer patterns than expected were generated for the years 2009 and 2010. It occurred due to the missing values of attributes to be considered for the generation of ARs.

5.4.2. Divergence in the hemoparasite attribute

The hemoparasite attribute of the dataset studied informs if the patient has the microfilaria and Chagas (genus *Trypanosoma*) disease. It was possible to notice a tendency for the appearance of microfilaria hemoparasites in the negative cases of malaria and the appearance of hemoparasites of the genus *Trypanosoma* in the positive cases. These results were previously unknown for the health specialists in the subject.

About 70% of the detected cases of the genus *Trypanosoma* are associated with Vivax. Malaria and Chagas disease, although transmitted by different vectors, both are prevalent in poor, remote, and riverine communities. In the patterns where the genus *Trypanosoma* appeared, only Marajo I and Tocantins health regions appear, both in Para. Some sources point to outbreaks of Chagas disease in some counties of these health regions, corroborating the validity of the results found [50,51].

Table 5
Support of the ARs for patients infected with *P. vivax* in the seven years of study.

Year	2009	2010	2011	2012	2013	2014	2015
Support	.710	.815	.838	.743	.728	.615	.949

In the state of Para, specifically, there are many occurrences of Chagas disease. A common way to get Chagas disease is by eating foods contaminated with the *Trypanosoma* genus, mainly acai and sugar cane. The triatomines, vectors of the disease, find ideal conditions in these plantations for their development. Hence, they can be crushed and sold together with the product. The feces of the insect also contaminate food.

The results also show that two months are associated with the Marajo I region: March and April. In total, more than 500 cases of vivax malaria and Chagas disease occur together in these months. However, according to the epidemiological bulletin, there are more cases of Chagas disease in the months from August to November, coinciding with the acai fruit harvest [52].

For positive microfilaria results, nine ARs are generated when the malaria result is negative. These ARs contemplate about 34% of the microfilaria cases that appear in the dataset. The health regions shown in the patterns were Alto Solimoes, Regional Jurua, and Regional Purus. These three health regions are found in the state of Amazonas.

Other values associated with positive cases of filariasis and negative cases of malaria are the summer months (December to March in Brazil). In these months, malaria does not predominate. The year 2011 also appears for Alto Solimoes and Regional Jurua. The absence of symptoms is another value that appears in the patterns. This last value is quite interesting because it indicates that the patient may be diagnosed with another disease that has not yet caused the symptoms. In the case of filariasis, the early diagnosis is of particular importance since this disease impairs the lymphatic system. It may obstruct the lymphatic vessels, causing swelling and thickening of the skin in the area that could lead to irreversible damage.

5.4.3. Divergence in the occupation attribute

Considering ARs for the negative result, there is a predominance of the occupation “road construction” among all other occupations. Besides, the support of the ARs with this occupation is considerably higher than the support of the other occupations, suggesting that the negative result for malaria is related to the “road construction” occupation. Another observation made is that only in the ARs dataset for negative results does the active detection type appear more frequently than the passive detection type. Another observation made is that only in the ARs dataset for negative results does the active detection type appear more frequently than the passive detection type.

These observations were previously unknown but are coherent. The chance of obtaining negative results is expected when health professionals seek potential infected people in risk groups (active surveillance). More positive cases are expected when people notice symptoms and look for care (passive surveillance) [53]. Since road and dam construction activities are, in general, carried out by regular workers, malaria infection monitoring constitutes a requirement of employment contract, producing many negative results.

A chi-square test was performed to study the level of relationship between the occupation attribute (reduced to two values: road construction and others) and the detection type attribute. The test was performed on a subset of the negative results from the preprocessed dataset. The *p-value* computed is less than the significance level of 5% and, therefore, the attributes were correlated. The outcome of this analysis indicates that many of the active detection campaigns occur at road construction sites.

Another observation about the occupation of road construction is that the patterns found for negative results for malaria are in the Alto

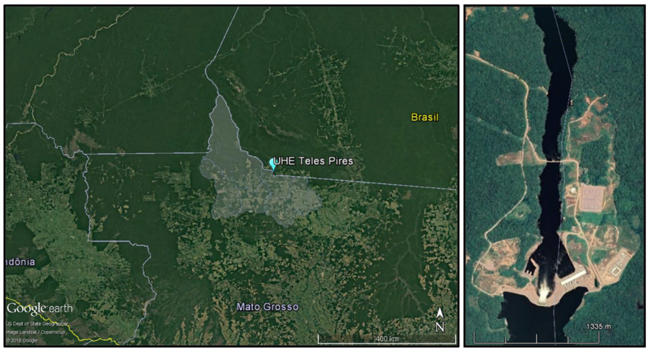

Fig. 1. Map with the location of the Alto Tapajos health region and the Teles Pires Hydroelectric Plant using Google maps [54].

Tapajos health region, in the state of Mato Grosso. The Teles Pires hydroelectric plant (UHE Teles Pires) is located in this region (on the border of the states of Para and Mato Grosso). It has an installed capacity of 1820 MW. It is the largest hydroelectric plant in the region, as depicted in Fig. 1.

Other values that compose these patterns are the year 2012, active detection type, and examination interval from 1 to 7 days. It is possible to infer that these examinations happened because of the Malaria Action and Control Plan, foreseen in the Basic Environmental Project of UHE Teles Pires. In this project, a commitment is made to conduct malaria examinations on employees of the dam construction on admission and dismissal, besides periodic examinations. Table 6 presents the five patterns with the highest support for the ARs generated in the subset of SivepMalaria for a negative result where the road construction occupation appears.

5.4.4. Divergence in the health regions

When analyzing the dataset of ARs for falciparum malaria type, it was curious to note that only three health regions appear in the ARs. The Codo health region in Maranhao appears in 21 of the 24 ARs generated. This health region corresponds to the counties Codo, Coroata, Sao Mateus do Maranhao, Alto Alegre do Maranhao, and Timbiras.

The Codo health region is not one of the regions with the highest malaria incidence rates (more precisely, it is 47th in a list of health regions in decreasing order of malaria cases). However, this health region stands out for having more cases of malaria caused by falciparum than by vivax. The year 2009 follows all the patterns of the Codo health region with the result of falciparum malaria type, in addition to values showing low education level, female gender, the existence of symptoms, agriculture as the occupational activity, and notification months between March and July. Table 7 shows the five patterns with the highest support for the falciparum malaria type.

These 21 patterns for the Codo health region characterize more than 70% of the records of the region (considering the result of the examination for falciparum malaria). It means that the falciparum in this region was well characterized in 2009. This type of analysis can help determine risk groups and guide prevention campaigns. This result indicates that DARA can be useful in online analysis to shed light on patterns that could be related to diseases outbreaks.

Table 6
ARs with road construction occupation for RHS = Negative.

LHS	Address	Support
{occupation = Road Constr}	Alto Tapajos	5.90×10^{-5}
{occupation = Road Constr ∧ exam.type = Thick and thin blood smears}	Alto Tapajos	5.89×10^{-5}
{occupation = Road Constr ∧ race = Mixed race}	Alto Tapajos	3.34×10^{-5}
{occupation = Road Constr ∧ exam.interval = 1 to 7 days}	Alto Tapajos	3.14×10^{-5}
{occupation = Road Constr ∧ notification.year = 2012}	Alto Tapajos	2.54×10^{-5}

Table 7
ARs for RHS = Falciparum.

LHS	Address	Support
{notification.year = 2009}	Codo	2.38×10^{-5}
{notification.year = 2009 ∧ symptom = Yes}	Codo	2.37×10^{-5}
{occupation = Agriculture ∧ notification.year = 2009}	Codo	1.74×10^{-5}
{notification.year = 2009 ∧ gender = Female}	Codo	8.94×10^{-6}
{exam.interval = 1 to 7 days ∧ notification.year = 2009}	Codo	8.75×10^{-6}

The other two health regions that appear are Marajo I (two patterns) and Triangulo (one pattern) located, respectively, in the states of Para and Amazonas. These two health regions have a much higher number of malaria cases when compared to Codo. However, these patterns do not cover even 2% of the total of these cases. Hence, the characterization of falciparum malaria type in these regions is not as understood as in Codo.

Additionally, in all the patterns where the time attributes (notification, examination, and treatment) appear, the values show that these times are short, that is, the notification, examination, and treatment are done fast. It may be related to the fact that falciparum malaria type is considered a medical emergency. Its treatment commonly starts within the first 24 h of the beginning of the fever. Since this type of malaria is considered the most aggressive type, it can cause many other medical problems if not treated in a short period.

5.4.5. Divergence in the race attribute

The DARA indicated that in the subsets of vivax malaria type and negative result, in the attribute “race”, the indigenous race contains fewer ARs than expected. Through the search of a probable reason for this fact, it was found that the indigenous race has a much lower percentage of passive detection than the active, compared to other races. It means that individuals of the indigenous race are less likely to look for a health unit and are more dependent on active detection campaigns.

The active detection type occurs mainly from campaigns where health professionals go to a specific location and subject individuals to the examination for malaria. Thus, as this is a health action, with routines and protocols, it has a better-defined pattern. In comparison, in the passive detection type, individuals seek care, and the variables involved tend to be more dispersed, leading to a lower number of patterns.

Considering the support of ARs for each race, it is observed that the indigenous population is more vulnerable to malaria. This information, coupled with the fact that this population tends not to seek medical care, is relevant. It may indicate that more active detection campaigns should occur in the locations where these individuals are, aiming the promotion of health for this population at risk.

5.5. Discussion

Despite the robust discovery made using DARA, this subsection discusses other computed divergences that did not produce new knowledge. Some considerations are made with these results, seeking to present the characteristics involved in the divergence. These examples show that values indicated as divergent do not guarantee interesting analysis.

The tables presented in Appendix B show the DARA results for all attributes in the four datasets for the different examination results. The values marked with ∇ or Δ are the values considered divergent since they are extreme (high or low) when compared to the other values of the attribute.

Divergences were also found in the attribute related to previous treatment in three parts: non-falciparum, vivax, and falciparum. For vivax and non-falciparum malaria type, the divergence indicates that more ARs have been generated for previous treatment in vivax. Similarly, in the falciparum part, the divergence indicates more ARs generated for previous treatment for the same illness. It may be correlated with the fact that when the patient is infected with malaria more than once, the type of malaria is usually recurrent. It is common even for the conditions/regions of infection to which that individual is regularly exposed.

For the education level attribute, there was an indication of divergence in the set of ARs for the vivax malaria type, where "illiterate" appears in fewer ARs than expected. About 67% of individuals who claimed to be illiterate and were diagnosed with vivax malaria type are included in one of the 199 ARs created. It indicates that lower education levels are better characterized in positive cases of malaria (vivax) since most individuals can be described in fewer ARs.

Some values showed divergence indexes, but they do not constitute an important divergence and, therefore, were not discussed here. Also, the attributes that have only one possible value cannot be analyzed by the DARA. In these cases, the attributes and their values do not appear in the Tables of Appendix B.

6. Conclusions

This paper proposes a novel approach targeting to filter interesting ARs, called DARA. It is built upon the hypothesis of attributes independence. It focuses on filtering ARs that contain at least one item whose frequency in the dataset is not in agreement with the number of ARs containing it.

DARA was evaluated in a real-world study of malaria occurrence in Brazilian Legal Amazon. It was possible to gather relevant information about malaria that was not clear during exploratory analysis. By using DARA, it was possible to obtain novel findings regarding malaria in Brazilian Legal Amazon. They would perhaps be hidden in a large amount of discovered ARs using traditional approaches.

DARA has two main advantages. The first one is that it targets DARs. They might be interesting from construction, as they are not in agreement with the item frequency in the dataset. Thus, they might not be easily discovered during exploratory data analysis. Using DARA, it was possible to observe divergence for the year of notification, type of hemoparasites, type of occupation, and the health regions in which malaria occurred, and the race of patients. Some of these divergent items were able to capture both problems regarding the quality of the dataset (missing values for years between 2009 and 2010) and novel findings regarding the behavior of malaria cases.

The second advantage is that it drives the analysis of a set of ARs containing divergent items. Such a process focuses the analyst's attention on tackling a subset of ARs. For example, when analyzing the hemoparasites divergent items, it was possible to shed light on several ARs regarding the attribute. It was possible to observe outbreak notification cases of malaria found at Marajo in March and April of 2011.

DARA was designed to find patterns in multi-dimensional datasets targeting a consequent attribute. As a limitation of the approach, it cannot be applied when users do not have a target question that leads to a consequent. Besides, although DARA cannot guarantee to lead to the discovery of interesting ARs (nor ARs of all types), it is a novel approach that targets their discovery. It can be applied in conjunction with other well-known approaches to filtering interesting ARs and direct their study. As future works, we intend to study and evaluate DARA used as standalone, prior, and after other traditional methods over multiple datasets.

Author contributions

All authors contributed equally to the study. Lais Baroni and Eduardo Ogasawara focused on the Methodology and Investigation. Rebecca Salles focused on the validation. Christovam Barcellos and Marcel Pedroso focused on the conceptualization and data curation. Samella Salles focused on the Writing – original draft. Eduardo Bezerra, Gustavo Guedes, and Fabio Porto focused on Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

See Table 8.

Table 8
Attributes of the preprocessed dataset.

Attribute	Definition	Value example
address	place of residence of the patient	Alto Tapajos
exam.interval	time interval between notification and examination	8 to 30 days
treatment.interval	time interval between examination and beginning of treatment	1 to 7 days
notification.interval	time interval between symptom and notification	on the same day
notification.month	month in which the notification was recorded	10
notification.year	year in which the notification was recorded	2010
migration	health region of residence different than that of notification	yes
autochthonous.case	health region of infection identical to that of residence	no
exam.result	result of examination	vivax
detection.type	type of detection	active
exam.type	type of examination	rapid diagnostic test
symptom	indicates if the patient felt a symptom	yes
hemiparasite	the result of the examination for other hemiparasites	microfilaria
previous.treatment	previous treatment for *P. vivax* or *P. falciparum*	no
occupation	main activity in the last 15 days	agriculture
education.level	level of education of the patient	illiterate
age	interval of the age of the patient	01 to 04 years old
race	race/color of the patient	mixed race
gender	gender of the patient	female

Appendix B

The following tables correspond to all attributes and values. It indicates the number of ARs that each attribute contains (in the 'ARs' column) and the number of records in which this attribute appears in the corresponding preprocessed subset (in the column 'Sup.'). Column 'L' shows the accumulated value in the *L* counter. Divergent values, according to the authors' interpretation, are marked with ∇ or Δ corresponding, respectively, to cases that generated fewer or more ARs than expected according to the hypothesis of attributes independence.

The following pages present the attributes of the generated ARs for all the analyzed attributes[5]. The corresponding subset is pointed at the top of the tables, being presented in order: negative, vivax, falciparum, and non-falciparum.

B.1. Negative

occupation

value	Sup.	ARs	L
agriculture	10,210	2	−1
domestic worker	5738	1	−1
hunting and fishing	1264	0	0
vegetal exploitation	237	0	0
panning	1,481	1	0
mining	72	0	0
grazing	478	0	0
tourism	935	0	0
traveler	551	0	0
road construction	1,843	10	Δ 2

detection.type

value	Sup.	ARs	L
active	8,264,503	16	0
passive	5,755,158	3	0

symptom

value	Sup.	ARs	L
yes	121,940	2	∇ −1
no	14,495	45	Δ 1

exam.type

value	Sup.	ARs	L
thick and thin blood smears	7,133,206	10	0
rapid diagnostic test	63,794	0	0

notification.interval

value	Sup.	ARs	L
1 to 7 days	94,762	3	−1
8 to 30 days	5,469	0	0
more than a month	6,164	0	0
on the same day	14,739	4	1

notification.month

value	Sup.	ARs	L
3	1,294,038	1	−5
6	1,226,022	1	−4
7	1,267,780	1	−4
5	1,289,089	4	0
9	1,104,465	1	0
10	1,092,004	0	0
11	1,048,590	0	0
12	918,865	0	0

[5] For attribute **address**, only divergent values are in tables.

8	1,221,692	3	1
1	1,209,200	4	3
2	1,144,301	2	3
4	1203615	5	Δ 6

notification.year

value	Sup.	ARs	L
2009	2,310,303	0	∇ −5
2010	2,377,425	0	∇ −5
2012	2,203,707	4	−1
2011	2,274,502	7	0
2014	1,587,803	5	2
2013	1,846,479	11	4
2015	1,419,442	11	Δ 5

hemiparasite

value	Sup.	ARs	L
microfilaria	8,706	9	0
trypanosoma	128	0	0
trypanosoma and microfilaria	10	0	0

education.level

value	Sup.	ARs	L
incomplete college	17,215	0	∇ −4
incomplete primary school	10,313	0	−3
illiterate	6,551	0	−3
complete secondary school	28,815	1	−2
complete college degree	264	0	0
incomplete secondary school	11,024	2	1
complete primary school	2,708	1	2
incomplete high school	1,586	1	3
complete high school	1,682	4	Δ 6

exam.interval

value	Sup.	ARs	L
1 to 7 days	72,424	14	0
8 to 30 days	3,523	0	0
more than a month	1,574	0	0

race

value	Sup.	ARs	L
indigenous	5,051	0	∇ −2
asian	383	0	0
mixed race	13,503	16	0
white	2,388	2	1
black	1,428	1	1

age

value	Sup.	ARs	L
45 to 54 years old	18,193	0	−4
35 to 44 years old	23,849	1	−2
15 to 24 years old	25,940	2	−1
25 to 34 years old	31,720	2	−1
55 to 64 years old	11,143	1	0
over than 75 years old	1,912	0	0
less than 01 year old	955	0	0
01 to 04 years old	4,321	1	1
65 to 74 years old	4,917	2	3
05 to 14 years old	12,702	5	4

gender

value	Sup.	ARs	L
male	82,235	2	∇ −1
female	53,401	6	Δ 1

address

value	Sup.	ARs	L
Entorno Manaus e Alto Rio Negro	107,158	0	−10
Noroeste Matogrossense	900	0	−4
Alto Solimoes	4,709	6	−3
Regional Purus	2,692	1	−3
Rio Negro e Solimoes	530	0	−3
Triangulo	476	0	−3
area Norte	445	0	−2
Jurua e Tarauaca/Envira	372	0	−2
Regional Jurua	1,160	2	−1
Ze Doca	7,485	43	1
Alto Tapajos	3,077	35	2
Presidente Dutra	475	1	2
Rosario	1,061	13	4
Cone Sul	349	9	11
Medio Norte Matogrossense	339	7	11

B.2. Vivax

occupation

value	Sup.	ARs	L
panning	64,811	26	∇ −6
agriculture	297,810	419	0
domestic worker	150,006	390	0
vegetal exploitation	20,518	78	0
mining	2,851	3	0
hunting and fishing	36,542	195	1
road construction	8,123	39	1
grazing	7,513	31	1
tourism	13,094	70	1
travaler	17,019	114	2

detection.type

value	Sup.	ARs	L
active	371,770	349	0
passive	1,112,483	629	0

symptom

value	Sup.	ARs	L
no	77,701	86	0
yes	1,406,552	800	0

exam.type

value	Sup.	ARs	L
thick and thin blood smears	783,552	794	0
rapid diagnostic test	3,396	7	0

notification.interval

value	Sup.	ARs	L
1 to 7 days	1,076,353	813	0
8 to 30 days	53,522	119	0
more than a month	11,495	43	0
on the same day	264,839	409	0

notification.month

value	Sup.	ARs	L
1	119,474	215	−3
7	155,674	333	−2
9	132,862	327	−1
11	118,054	238	−1
2	105,822	215	0
10	122,713	293	0
12	100,647	212	0
3	106,978	248	1
5	127,719	331	1
6	140,487	336	1
8	146,979	370	1
4	106,844	292	3

notification.year

value	Sup.	ARs	L
2009	257,695	279	∇ −5
2010	282,587	325	∇ −4
2013	168,211	327	1
2014	138,203	300	1
2011	257,159	433	2
2012	234,038	421	2
2015	146,360	365	3

hemiparasite

value	Sup.	ARs	L
microfilaria	2,530	2	∇ −1
trypanosoma and microfilaria	1,105	0	0
trypanosoma	1,802	11	Δ 1

education.level

value	Sup.	ARs	L
illiterate	163,252	199	∇ −4
incomplete primary school	395,831	309	−3
complete high school	54,104	242	0
complete college degree	5,755	47	0
incomplete college	28,396	113	0
incomplete secondary school	290,435	433	1
complete primary school	97,676	320	2
complete secondary school	130,773	343	2
incomplete high school	53,512	252	2

exam.interval

value	Sup.	ARs	L
1 to 7 days	72,424	337	0
8 to 30 days	3,523	23	0
more than a month	1,574	0	0

race

value	Sup.	ARs	L
indigenous	110,863	208	∇ −2
asian	10,671	72	0
mixed race	553,747	625	0
white	67,850	341	1
black	43,773	240	1

age

value	Sup.	ARs	L
35 to 44 years old	173,387	178	−2
25 to 34 years old	261,772	246	−1
45 to 54 years old	107,019	118	−1
55 to 64 years old	53,522	94	−1
05 to 14 years old	361,347	566	0
15 to 24 years old	317,632	528	0
65 to 74 years old	21,220	69	0
over 75 years old	8,893	31	0
01 to 04 years old	155,625	451	2
less than 1 year old	23,835	182	3

gender

value	Sup.	ARs	L
male	907,669	485	∇ −1
female	576,468	536	Δ 1

previous.treatment

value	Sup.	ARs	L
falciparum	9,520	2	∇ −1
vivax	116,611	494	0
vivax and falciparum	6,614	12	Δ 1

treatment.interval

value	Sup.	ARs	L
1 to 7 days	27,540	156	0
8 to 30 days	1,046	0	0
more than a month	2,941	15	0
treatment before examination	7,878	56	0

address

value	Sup.	ARs	L
Tapajos	72,202	108	−26
Jurua e Tarauaca/Envira	162,994	201	−23
Alto Tapajos	1,041	0	−13
Pedreiras	1,167	1	−12
Pais Fronteira	4,579	32	−9
Presidente Dutra	783	0	−9
Regional Jurua	86,782	376	−7
Centro Norte	54,970	358	−6
Ze Doca	5,631	52	−6
area Norte	30,322	270	−5
Alto Solimoes	84,758	398	−4
Marajo II	124,703	415	−4
Santa Ines	1,460	13	−4
Rosario	368	0	−3
Chapadinha	224	0	−2
Viana	230	0	−2
Bico do Papagaio	197	0	−1
Caxias	167	0	−1
Entorno Manaus e Alto Rio Negro	175,978	498	−1
Itapecuru Mirim	175	0	−1
Rio Caetes	6,530	175	−1
Vale do Peixoto	1,021	11	−1
Acailandia	585	6	1
Bacabal	766	8	1
Cafe	2,488	87	1
Imperatriz	697	6	1
Medio Amazonas	6,055	180	1
Noroeste Matogrossense	6,636	186	1
Regional Purus	50,375	401	1
Zona da Mata	1,213	40	1
Baixo Acre e Purus	8,065	252	2
Cone Sul	943	28	2

Metropolitana I	4,356	174	2
Vale do Guapore	1,653	58	2
Araguaia	2,977	151	3
Baixo Amazonas	20,085	299	3
Madeira-Mamore	124,415	589	3
Marajo I	19,803	284	3
Central	7,979	256	4
Medio Norte Matogrossense	204	1	4
Metropolitana II	2,345	136	4
Metropolitana III	21,743	337	4
Rio Negro e Solimoes	28,470	373	4
Sul	22,018	340	4
area Sudoeste	24,329	366	5
Codo	355	5	5
Sao Luis	865	35	5
Pinheiro	4,131	181	6
area Central	41,970	433	7
Baixada Cuiabana	537	11	7
Sudoeste Matogrossense	452	10	7
Lago de Tucurui	33,794	416	8
Teles Pires	816	41	8
Xingu	39,613	442	9
Alto Acre	165	2	10
Vale do Jamari	33,626	452	12

B.3. Falciparum

occupation

value	Sup.	ARs	L
agriculture	32,408	1	0
hunting and fishing	3,922	0	0
road construction	531	0	0
domestic worker	18,833	0	0
vegetal exploitation	4,710	0	0
panning	28,897	0	0
mining	682	0	0
grazing	1,014	0	0
tourism	759	0	0
travaler	1,683	0	0

detection.type

value	Sup.	ARs	L
passive	142,254	0	∇ −1
active	50,030	1	Δ 1

symptom

value	Sup.	ARs	L
yes	12,196	1	0
no	180,088	1	0

notification.interval

value	Sup.	ARs	L
1 to 7 days	134,516	1	0
8 to 30 days	7,313	0	0
more than a month	1,328	0	0
on the same day	36,893	1	0

notification.month

value	Sup.	ARs	L
7	18,553	0	−4
8	17,445	0	−4
5	15,410	0	−3
6	16,439	0	−3
9	16,150	0	−3

10	15,993	0	−3
11	17,189	0	−3
12	14,850	0	−2
1	17,433	3	2
2	14,944	2	Δ 7
3	14,079	2	Δ 8
4	13,799	2	Δ 8

notification.year

value	Sup.	ARs	L
2010	40,692	0	∇ −3
2012	27,950	0	−2
2013	24,302	0	−1
2014	18,144	0	−1
2009	40,662	15	1
2011	27,722	1	1
2015	12,812	2	Δ 5

education.level

value	Sup.	ARs	L
incomplete primary school	58,432	0	−2
complete primary school	12,439	0	0
complete secondary school	14,588	0	0
complete high school	5,994	0	0
incomplete high school	6,019	0	0
complete college degree	596	0	0
incomplete college	2,477	0	0
incomplete secondary school	40,893	2	1
illiterate	24,346	2	1

exam.interval

value	Sup.	ARs	L
1 to 7 days	7,883	3	0
8 to 30 days	464	0	0
more than a month	236	0	0

age

value	Sup.	ARs	L
15 to 24 years old	40,634	0	−2
25 to 34 years old	38,034	0	−1
01 to 04 years old	13,908	0	0
45 to 54 years old	18,185	0	0
55 to 64 years old	7,966	0	0
65 to 74 years old	2,819	0	0
over 75 years old	1,043	0	0
less than 1 year old	1,905	0	0
05 to 14 years old	40,227	2	1
35 to 44 years old	27,563	1	2

gender

value	Sup.	ARs	L
male	119,998	0	∇ −1
female	72,263	1	Δ 1

previous.treatment

value	Sup.	ARs	L
vivax	9,236	0	∇ −1
vivax and falciparum	1,023	0	0
falciparum	3,417	1	Δ 1

treatment.interval

value	Sup.	ARs	L
1 to 7 days	4,340	2	0
8 to 30 days	167	0	0
more than a month	388	0	0
treatment before examination	937	0	0

address

value	Sup.	ARs	L
Alto Solimoes	10,113	0	−3
area Central	5,435	0	−3
area Norte	6,880	0	−3
Centro Norte	6,658	0	−3
Entorno Manaus e Alto Rio Negro	9,302	0	−3
Jurua e Tarauaca/Envira	34,873	0	−3
Madeira-Mamore	7,117	0	−3
Marajo II	21,487	0	−3
Regional Jurua	9,903	0	−3
Tapajos	29,494	0	−3
Lago de Tucurui	4,470	0	−2
Regional Purus	4,515	0	−2
Vale do Jamari	4,198	0	−2
area Sudoeste	3,244	0	−1
Baixo Amazonas	3,654	0	−1
Carajas	603	0	−1
Metropolitana III	3,170	0	−1
Noroeste Matogrossense	856	0	−1
Pais Fronteira	959	0	−1
Rio Caetes	606	0	−1
Rio Madeira	1,799	0	−1
Rio Negro e Solimoes	1,930	0	−1
Sul	1,579	0	−1
Tocantins	2,028	0	−1
Xingu	3,066	0	−1
Triangulo	4,996	1	8
Marajo I	3,766	2	13
Codo	532	21	Δ 27

B.4. Non-Falciparum

occupation

value	Sup.	ARs	L
panning	277	0	−1
agriculture	2,483	9	0
hunting and fishing	701	6	0
road construction	7	0	0
domestic worker	961	6	0
mining	10	0	0
grazing	23	0	0
tourism	15	0	0
travaler	73	0	0
vegetal exploitation	74	3	1

detection.type

value	Sup.	ARs	L
active	5,090	3	∇ −1
passive	4,968	4	Δ 1

symptom

value	Sup.	ARs	L
no	498	1	0
yes	9,560	4	0

notification.interval

value	Sup.	ARs	L
1 to 7 days	6,090	4	−1
8 to 30 days	326	2	0
more than a month	91	0	0
on the same day	3,048	5	1

notification.month

value	Sup.	ARs	L
5	1,027	4	−3
7	1,125	5	−2
11	687	4	−2
9	1,228	7	−1
10	1,062	5	−1
1	551	4	0
2	478	2	0
8	1,114	6	0
12	478	2	0
3	621	5	2
4	665	5	2
6	1,022	8	5

notification.year

value	Sup.	ARs	L
2015	1,658	2	∇ −2
2012	4,226	8	0
2013	3,080	6	0
2011	397	3	1
2014	697	3	1

education.level

value	Sup.	ARs	L
incomplete secondary school	1,173	2	−2
complete secondary school	332	0	−1
complete high school	261	0	−1
incomplete high school	308	0	−1
incomplete primary school	3,527	7	0
incomplete college	20	0	0
complete primary school	742	3	1
illiterate	1,164	6	1
complete college Degree	42	1	3

exam.interval

value	Sup.	ARs	L
1 to 7 days	341	3	0
8 to 30 days	76	0	0
more than a month	12	0	0

race

value	Sup.	ARs	L
mixed race	4,649	1	−1
asian	82	1	0
indigenous	4,657	7	0
black	230	1	0
white	433	5	1

age

value	Sup.	ARs	L
35 to 44 years old	746	0	−3
15 to 24 years old	1,908	3	−2
25 to 34 years old	1,262	3	−1
05 to 14 years old	3,289	8	0
45 to 54 years old	425	3	0
65 to 74 years old	128	0	0
over 75 years old	64	0	0
01 to 04 years old	1,687	7	1
55 to 64 years old	213	3	1
less than 1 year old	336	4	4

gender

value	Sup.	ARs	L
male	5,543	2	∇ −1
female	4,515	5	Δ 1

previous.treatment

value	Sup.	ARs	L
falciparum	72	0	∇ −1
vivax	1,453	10	0
vivax and falciparum	71	1	Δ 1

treatment.interval

value	Sup.	ARs	L
1 to 7 days	85	1	0
8 to 30 days	16	0	0
more than a month	52	0	0
treatment before examination	112	2	0

address

value	Sup.	ARs	L
Alto Solimoes	960	3	−16
Entorno Manaus e Alto Rio Negro	837	4	−11
Regional Purus	564	4	−9
Marajo II	1,811	13	−8
Tapajos	690	5	−8
Tocantins	933	14	−4
Jurua e Tarauaca/Envira	409	6	−3
area Sudoeste	28	0	−2
Regional Jurua	497	10	−1
Vale do Jamari	57	1	−1
Xingu	1,006	19	−1
Medio Amazonas	18	1	1
Madeira-Mamore	121	10	3
Marajo I	102	9	3
Rio Negro e Solimoes	27	4	3
area Norte	760	20	4
Lago de Tucurui	66	8	5
area Central	319	18	7
Baixo Amazonas	490	23	9
Metropolitana III	125	19	9
Araguaia	69	16	10
Metropolitana II	88	17	10

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2020.103512.

References

[1] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3 ed., Morgan Kaufmann, Haryana, India; Burlington, MA, 2011.

[2] K. Lodhi, Survey on frequent pattern mining, Int. J. Eng. Sci. Mathe. 2 (2013) 64.

[3] P.-N. Tan, V. Kumar, Interestingness measures for association patterns: A perspective, in: Proc. of Workshop on Postprocessing in Machine Learning and Data Mining, 2000, pp. 00–036.

[4] H. Zhang, B. Padmanabhan, A. Tuzhilin, On the discovery of significant statistical quantitative rules, in, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 374–383.

[5] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Discovering frequent closed itemsets for association rules, in: International Conference on Database Theory, Springer, 1998, pp. 398–416.

[6] R.J. Bayardo, R. Agrawal, D. Gunopulos, Constraint-based rule mining in large, dense databases, Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337), IEEE, 1999, pp. 188–197.

[7] G. Dong, J. Li, Efficient Mining of Emerging Patterns: Discovering Trends and Differences, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '99, ACM, New York, NY, USA, 1999, pp. 43–52. URL http://doi.acm.org/10.1145/312129.312191. doi: 10.1145/312129.312191.

[8] B. Liu, W. Hsu, L.-F. Mun, H.-Y. Lee, Finding interesting patterns using user expectations, IEEE Trans. Knowl. Data Eng. 11 (1999) 817–832.

[9] A. Silberschatz, A. Tuzhilin, On subjective measures of interestingness in knowledge discovery., in: KDD, vol. 95, 1995, pp. 275–281.

[10] S. Sahar, Interestingness via what is not interesting, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 1999, pp. 332–336.

[11] L. Gadár, J. Abonyi, Frequent pattern mining in multidimensional organizational networks, Sci. Rep. 9 (2019).

[12] K. Mcgarry, A survey of interestingness measures for knowledge discovery, Knowledge Eng. Rev. 20 (2005) 39–61.

[13] C. Bressan, P. Brasil, Malária, 2013. URL https://agencia.fiocruz.br/malaria.

[14] WHO, Key malaria facts, 2018. URL http://www.who.int/en/news-room/fact-sheets/detail/malaria.

[15] WHO, World malaria report 2017 - World Health Organization, 2017.

[16] U. Confalonieri, C. Margonari, A. Quintão, Environmental change and the dynamics of parasitic diseases in the Amazon, Acta Trop. 129 (2014) 33–41.

[17] J. Sachs, P. Malaney, The economic and social burden of malaria, Nature 415 (2002) 680–685.

[18] Y. Aumann, Y. Lindell, A statistical theory for quantitative association rules, in, Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 1999, pp. 261–270.

[19] H. Toivonen, others, Sampling large databases for association rules, in: VLDB, vol. 96, 1996, pp. 134–145.

[20] R. Agrawal, R. Srikant, others, Fast algorithms for mining association rules, in: Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, 1994, pp. 487–499.

[21] J. Han, J. Pei, Y. Yin, Mining Frequent Patterns Without Candidate Generation, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD '00, ACM, New York, NY, USA, 2000, pp. 1–12.

[22] M.J. Zaki, Scalable algorithms for association mining, IEEE Trans. Knowl. Data Eng. 12 (2000) 372–390.

[23] F. Berzal, I. Blanco, M. Vila, others, Measuring the accuracy and interest of association rules: A new framework, Intell. Data Anal. 6 (2002) 221–235.

[24] R.J. Larsen, M.L. Marx, An Introduction to Mathematical Statistics and Its Applications, 4th ed., Prentice Hall, Upper Saddle River, N.J., 2005.

[25] Z. Zheng, R. Kohavi, L. Mason, Real world performance of association rule algorithms, in, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 401–406.

[26] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

[27] V. Dhar, Data science and prediction, Commun. ACM 56 (2013) 64–73.

[28] B. Liu, W. Hsu, S. Chen, Y. Ma, Analyzing the subjective interestingness of association rules, IEEE Intell. Syst. Appl. 15 (2000) 47–55.

[29] R. Srikant, Q. Vu, R. Agrawal, Mining association rules with item constraints., in: Kdd, vol. 97, 1997, pp. 67–73.

[30] P.-N. Tan, V. Kumar, J. Srivastava, Selecting the right objective measure for association analysis, Informat. Syst. 29 (2004) 293–313.

[31] S. Brin, R. Motwani, J.D. Ullman, S. Tsur, Dynamic itemset counting and implication rules for market basket data, Acm Sigmod Rec. 26 (1997) 255–264.

[32] M. Hahsler, K. Hornik, New probabilistic interest measures for association rules, Intell. Data Anal. 11 (2007) 437–455.

[33] T. Wu, Y. Chen, J. Han, Re-examination of interestingness measures in pattern mining: a unified framework, Data Min. Knowl. Discov. 21 (2010) 371–397.

[34] P.-N. Tan, V. Kumar, J. Srivastava, Selecting the right interestingness measure for association patterns, in, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 32–41.

[35] R.J. Bayardo Jr, Efficiently mining long patterns from databases, in: ACM Sigmod Record, vol. 27, ACM, 1998, pp. 85–93.

[36] B. Padmanabhan, A. Tuzhilin, A Belief-Driven Method for Discovering Unexpected Patterns., in: KDD, vol. 98, 1998, pp. 94–100.

[37] R.T. Ng, L.V. Lakshmanan, J. Han, A. Pang, Exploratory mining and pruning optimizations of constrained associations rules, in: ACM Sigmod Record, vol. 27, ACM, 1998, pp. 13–24.

[38] W. Gan, J.C.-W. Lin, H.-C. Chao, H. Fujita, S.Y. Philip, Correlated utility-based pattern mining, Inf. Sci. 504 (2019) 470–486.

[39] P. Fournier-Viger, Y. Zhang, J.C.-W. Lin, H. Fujita, Y.S. Koh, Mining local and peak high utility itemsets, Inf. Sci. 481 (2019) 344–367.

[40] P. Fournier-Viger, C.-W. Wu, S. Zida, V.S. Tseng, FHM: Faster high-utility itemset mining using estimated utility co-occurrence pruning, in: International Symposium on Methodologies for Intelligent Systems, Springer, 2014, pp. 83–92.

[41] P. Fournier-Viger, J.C.-W. Lin, T. Dinh, H.B. Le, Mining correlated high-utility itemsets using the bond measure, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, 2016, pp. 53–65.

[42] A. Soulet, C. Raïssi, M. Plantevit, B. Cremilleux, Mining dominant patterns in the sky, 2011 IEEE 11th International Conference on Data Mining, IEEE, 2011, pp. 655–664.

[43] Y. Yan, L. Cao, S. Madden, E.A. Rundensteiner, SWIFT: mining representative patterns from large event streams, Proc. VLDB Endowment 12 (2018) 265–277.

[44] L. Pellegrina, F. Vandin, Efficient mining of the most significant patterns with permutation testing, in, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 2070–2079.

[45] A. Wiefels, B. Wolfarth-Couto, N. Filizola, L. Durieux, M. Mangeas, Accuracy of the malaria epidemiological surveillance system data in the state of Amazonas, Acta Amazonica 46 (2016) 383–390.

[46] M. Hahsler, C. Buchta, B. Gruen, K. Hornik, arules: Mining Association Rules and Frequent Itemsets, 2018. URL https://CRAN.R-project.org/package=arules.

[47] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

[48] A. Sternberg, D. Carvalho, L. Murta, J. Soares, E. Ogasawara, An analysis of Brazilian flight delays based on frequent patterns, Transport. Res. Part E: Logist. Transport. Rev. 95 (2016) 282–298.

[49] L. Baroni, M. Pedroso, C. Barcellos, R. Salles, S. Salles, B. Paixão, A. Chrispino, G. Guedes, E. Ogasawara, An Integrated Dataset of Malaria Notifications in the Legal Amazon, Technical Report, https://doi.org/10.7303/syn21552203, 2020.

[50] A.L.R. Roque, S.C. Xavier, M. Gerhardt, M.F. Silva, V.S. Lima, P.S. D'Andrea, A.M. Jansen, Trypanosoma cruzi among wild and domestic mammals in different areas of the abaetetuba municipality (pará state, Brazil), an endemic chagas disease transmission area, Veterinary Parasitol. 193 (2013) 71–77.

[51] A.Y. das Neves Pinto, S.A. Valente, V. da Costa Valente, J.A.G. Ferreira, J.R. Coura, Acute phase of chagas disease in the brazilian amazon region: study of 233 cases from pará, amapá and maranhão observed between 1988 and 2005, Revista da Sociedade Brasileira de Medicina Tropical 41 (2008).

[52] K.S. Pereira, F.L. Schmidt, R.L. Barbosa, A.M. Guaraldo, R.M. Franco, V.L. Dias, L.A. Passos, Transmission of chagas disease (american trypanosomiasis) by food, in: Advances in Food and Nutrition Research, vol. 59, Elsevier, 2010, pp. 63–85.

[53] W.H. Organization, et al., Disease surveillance for malaria control: an operational manual, 2012.

[54] Google, Google maps, Technical Report, https://www.google.com.br/maps, 2020.