UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Mineração de regras temporais multivariadas aplicada ao comércio internacional

Eliane Gniech Karasawa

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Eliane Gniech Karasawa

Mineração de regras temporais multivariadas aplicada ao comércio internacional

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestra em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Elaine Parros Machado de Sousa

USP – São Carlos
Julho de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

K18m

Karasawa, Eliane Gniech
Mineração de regras temporais multivariadas aplicada ao comércio internacional / Eliane Gniech Karasawa; orientadora Elaine Parros Machado de Sousa. -- São Carlos, 2024.
122 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Comércio Internacional. 2. Mineração de Regras Temporais Multivariadas. 3. Série Temporal Multivariada. I. Sousa, Elaine Parros Machado de, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Eliane Gniech Karasawa

Multivariate rules mining applied to international trade

Dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Master in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Elaine Parros Machado de Sousa

USP – São Carlos
July 2024

Este trabalho é dedicado à minha família, que sempre me apoiou e incentivou.

AGRADECIMENTOS

Agradeço à minha família, pelo apoio incondicional e por estarem sempre presentes.

Aos meus amigos do GBDI, em especial ao Marcus Vinicius, ao Lucas Scabora e ao Afonso Matheus que contribuiram diretamente auxiliando nos experimentos, dando conselhos e ajudando a manter um mínimo de socialização mesmo em tempos de pandemia.

À minha orientadora Profª. Drª. Elaine Parros, por confiar no meu potencial e me acompanhar nessa jornada, sempre se esforçando ao máximo para me auxiliar. Não poderia ter melhor orientação.

À Universidade de São Paulo, por toda estrutura física e recursos fonecidos.

Aos professores e funcionários do ICMC-USP que contribuiram direta ou indiretamente para este trabalho.

Ao CNPq e CAPES pelo apoio financeiro à realização deste trabalho.

“To doubt everything, or, to believe everything, are two equally convenient solutions; both dispense with the necessity of reflection.”

(Henry Poincaré)

RESUMO

KARASAWA, E. G. **Mineração de regras temporais multivariadas aplicada ao comércio internacional**. 2024. 122 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

O comércio internacional influencia significativamente a economia mundial e a análise de seus dados utilizando ferramentas computacionais se mostra como uma alternativa para melhor compreensão da economia global, altamente complexa e interdependente. Os dados econômicos apresentam alta correlação com o período a que se referem, por exemplo, durante uma crise mundial espera-se que diversos países apresentem variação em seus índices, em geral redução na tendência de crescimento ou queda, independentemente da sua situação econômica. Por tanto, técnicas de análise e mineração de séries temporais que considerem a característica temporal mostram-se como uma abordagem promissora. O presente trabalho explorou a tarefa de mineração de regras temporais aplicada a dados do comércio internacional. Algoritmos existentes de mineração de regras temporais com mais que duas variáveis não tratam séries heterogêneas e incompletas nem retornam informação das ocorrências das regras nas séries. O método proposto, eTRUMiner é capaz de minerar múltiplas séries heterogêneas e incompletas utilizando estratégias para redução da complexidade temporal e espacial. Os resultados automatizados podem auxiliar o economista na tarefa de avaliação e compreensão das informações obtidas. As regras retornadas são compostas de duas ou mais variáveis distintas e permitem a localização de suas ocorrências nas séries. A aplicação do eTRUMiner sobre as séries do comércio internacional permite verificar comportamentos globais esperados, como o crescimento econômico nos países, e também características específicas como regras em períodos de crise.

Palavras-chave: Comércio internacional, Mineração de Regras Temporais Multivariadas, Série Temporal Multivariada.

ABSTRACT

KARASAWA, E. G. **Multivariate rules mining applied to international trade**. 2024. 122 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

International trade significantly influences the global economy, and using computational tools to analyze its data arises as an alternative for better understanding the highly complex and interdependent global economy. Economic data show a high correlation with the period they refer to; for instance, during a global crisis, it is expected that various countries will display variations in their indices, generally a reduction in growth trends or a decline, regardless of their economic situation. Therefore, techniques for analyzing and mining time series considering the temporal characteristic appear promising approaches. The present work explored the task of mining temporal rules from international trade data. Existing algorithms for mining temporal rules with more than two variables do not handle heterogeneous and incomplete series nor return information on the occurrences of the rules in the series. The proposed method, eTRUMiner, is capable of mining multiple heterogeneous and incomplete series using strategies to reduce temporal and spatial complexity. The automated results can assist economists in evaluating and understanding the obtained information. Returned rules are composed of two or more distinct variables and allow the location of their occurrences in the series. The application of eTRUMiner on international trade series allows the verification of expected global behaviors, such as economic growth in countries, and specific characteristics like rules during crisis periods.

Keywords: International Trade, Multivariate Temporal Rules Mining, Multivariate Time Series.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ALGORITMOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CEPII	*Centre d'Études Prospectives et d'Informations Internationales*
CLEARMiner	*CLimatE Association patteRns Miner*
ECI	*Economic Complexity Index*
eTRUMiner	*extended Temporal RUles Miner*
EXP	exportação
FMI	Fundo Monetário Internacional
IMP	importação
KDD	*Knowledge Discovery in Databases*
PAA	*Piecewise Aggregate Approximation*
PIB	Produto Interno Bruto
SAX	*Symbolic Aggregate Approximation*
TARM	*Temporal Association Rule Mining*

LISTA DE SÍMBOLOS

S — Conjunto de dados

N — Número de séries multivariadas no conjunto de dados S

s — Série temporal

n — Número de observações na série temporal s

δ — Número de variáveis da série s

obs_i^X — Observação da variável X no tempo t_i

A — Antecedente da regra

C — Consequente da regra

I — Lista de itens

sup — Suporte

$conf$ — Confiança

sup_{min} — Suporte mínimo

$conf_{min}$ — Confiança mínima

Δt — Característica temporal da regra temporal

var_X — Variável X

$s[var_X]$ — Série temporal univariada da variável X

$s'[var_X]$ — Série discretizada da série $s[var_X]$

L — Quantidade de elementos discretizados na série discretizada $s'[var_X]$

α_{t_i,t_f}^X — Elemento discretizado pertencente a série discretizada $s'[var_X]$ referente ao tempo inicial t_i e tempo final t_f da série temporal $s[var_X]$

α_i^X — i-ésimo elemento discretizado da série discretizada $s'[var_X]$

$[var_X, \alpha_i^X]$ — Padrão do elemento discretizado α_i^X na variável X

w — Janela temporal

T — Número de transações no conjunto de dados S

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A Descoberta de Conhecimento em Bases de Dados, do inglês *Knowledge Discovery in Databases* (KDD), é um processo interativo e iterativo empregado para identificar no conjunto de dados, padrões novos, válidos, úteis e compreensíveis (FAYYAD *et al.*, 1996). Cunhado na década de 1990 com o aumento no volume de dados sendo gerados, o termo destaca a finalidade da análise, que é a obtenção de conhecimento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O processo de KDD pode ser dividido em três etapas principais (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; HAN; PEI; KAMBER, 2011; REZENDE *et al.*, 2003), sumarizadas em Pré-Processamento, Mineração de Dados e Pós-Processamento. A aplicação de algoritmos visando a extração de padrões é realizada na etapa de mineração de dados, após o tratamento do conjunto no pré-processamento. Durante o pós-processamento, os padrões extraídos são selecionados com o uso de métricas de avaliação.

Dentre as tarefas de mineração de dados, a extração de regras de associação possui grande potencial por fornecer padrões com simplicidade e explicabilidade. Em particular, na mineração de regras temporais, além das relações de causalidade entre antecedente e consequente da regra (AGRAWAL; IMIELIŃSKI; SWAMI, 1993), busca-se uma compreensão mais aprofundada da ordem dos eventos e tempo de ocorrência (SEGURA-DELGADO *et al.*, 2020). Essa tarefa é comumente aplicada a séries temporais, que estão presentes em diversas áreas de aplicação como saúde, biologia, geografia e economia.

Mais especificamente na área econômica, as séries temporais podem representar dados como preços de ações, *commodities*, valores de importação e exportação, Produto Interno Bruto (PIB). Nesse contexto, o foco deste trabalho são os dados de comércio internacional, baseado em mercadorias e serviços transacionados entre países. O volume de dados gerados por essas transações tem crescido rapidamente (UNCTAD, 2019), atingindo atualmente a ordem de bilhões

de registros de dados. Logo, o armazenamento, a integração, a disseminação e principalmente a análise desses dados requer o auxílio de ferramentas computacionais.

1.1 Contextualização e Motivação

O comércio internacional influencia significativamente a economia mundial (EUROSTAT, 2006), direcionando tomadas de decisão tanto em setores públicos quanto privados. A análise de seus dados permite a compreensão da situação econômica e do estado de desenvolvimento dos países, assim como o entendimento do relacionamento econômico entre países e entre produtos. Portanto, suas estatísticas apresentam relevância em âmbito nacional e internacional.

De acordo com o Relatório de Comércio e Desenvolvimento (UNCTAD, 2023), as assimetrias existentes no comércio internacional têm aumentado na última década, principalmente após o período da pandemia por COVID-19. Verifica-se neste contexto, o crescimento na distribuição assimétrica do fluxo de transações e o aumento na concentração de exportações, no endividamento e na desigualdade do desenvolvimento tecnológico.

A análise de dados do comércio internacional auxiliada por ferramentas computacionais pode colaborar para melhor compreensão da economia global, altamente complexa e interdependente. Com o entendimento mais aprofundado dos perfis econômicos existentes, compreendendo relações entre características econômicas e entre nações, pode-se planejar estratégias de desenvolvimento e parcerias para melhorar relações econômicas e desenvolver a economia.

Dados sobre o comércio internacional são fornecidos gratuitamente por diversos órgãos, tais como o *Centre d'Études Prospectives et d'Informations Internationales* (CEPII) [1], o Observatório de Crescimento de Harvard [2] e o Fundo Monetário Internacional (FMI) [3]. Dentre os índices disponibilizados, destacam-se os valores de importação, exportação, Produto Interno Bruto (PIB) e Índice de Complexidade Econômica (ECI - do inglês *Economic Complexity Index*).

No conjunto de trabalhos recentes que exploram esses dados, Tacchella, Mazzilli e Pietronero (2018) buscam simplificar a predição do PIB com o auxílio do ECI, utilizando uma abordagem estatística. Resultados apontaram um aumento na precisão de mais de 25% na predição da janela temporal de 5 anos em relação às predições amplamente utilizadas fornecidas pelo FMI. Esse resultado indica alta correlação entre as medidas de PIB e ECI.

A influência do ECI em crises fiscais é avaliada no trabalho de Gomez-Gonzalez, Uribe e Valencia (2023) também com uma abordagem estatística. Através de uma análise de sobrevivência utilizando modelos de risco, compara-se diversos fatores econômicos como o PIB, a soma de importações e exportações e ECI. Conclui-se que o aumento em uma unidade no ECI reduz em

[1] CEPII <http://www.cepii.fr/CEPII/fr/cepii/cepii.asp>
[2] *Growth Lab* <https://growthlab.hks.harvard.edu/home>
[3] FMI <https://www.imf.org/en/Home>

até metade o risco de crises fiscais, reforçando a importância de considerar o ECI em análises macroeconômicas.

Em Dar *et al.* (2020), utiliza-se a combinação linear de observações anteriores associada a uma modelagem matemática para predizer índices do comércio internacional da Coreia do Sul. Dados de importação e exportação, PIB, PIB *per capita* e ECI são avaliados visando predizer o seu comportamento e a sua performance nos 6 anos seguintes. O conjunto de dados é similar ao utilizado neste trabalho de mestrado, porém específico para os dados da Coreia do Sul, com enfoque na predição e apresentando resultados com validação superficial.

Os dados econômicos apresentam alta correlação com o período a que se referem, por exemplo, durante uma crise mundial espera-se que diversos países apresentem variação em seus índices, em geral redução na tendência de crescimento ou queda, independentemente da sua situação econômica. Contudo, a análise desses dados é realizada com enfoque estatístico (TACCHELLA; MAZZILLI; PIETRONERO, 2018; GOMEZ-GONZALEZ; URIBE; VALENCIA, 2023; DAR *et al.*, 2020), sem avaliar relações de causalidade temporal existentes e que possuem grande potencial informativo. Dessa forma, técnicas de análise e mineração de séries temporais que considerem a característica temporal e identifiquem padrões úteis, como as regras temporais, mostram-se como uma abordagem promissora.

O conjunto de dados do comércio internacional compõe-se de dados com origem variada, cada índice sendo fornecido por um órgão distinto, com durações distintas e valores faltantes. Para a análise do comércio internacional, os índices econômicos podem ser tratados como variáveis de uma série temporal multivariada que refere-se a um país. O conjunto integrado constitui-se de séries denominadas “heterogêneas”, em que cada variável da série possui uma duração, e “incompletas”, contendo observações e variáveis faltantes. A mineração de regras temporais multivariadas a partir dessas séries pode auxiliar na compreensão do estado comercial e financeiro dos países, bem como países similares.

Para uma avaliação mais completa dos dados do comércio internacional, a abordagem utilizada deve ser capaz de lidar com conjuntos de séries heterogêneas incompletas, retornando regras compostas de duas ou mais variáveis. Além disso, a informação das ocorrências de cada regra nas séries temporais permite uma compreensão mais completa do período e países a que se refere. Contudo, não foi encontrado, durante essa pesquisa, nenhum trabalho que integre essas características, conforme discutido na Seção 2.3. Trabalhos relacionados, como o MOWCATL (HARMS; DEOGUN, 2004) e o CLEARMiner (ROMANI *et al.*, 2010) retornam regras com apenas 2 variáveis e o TARM minera regras com mais variáveis mas apenas de uma série temporal completa. As ocorrências das regras nas séries temporais são retornadas apenas no CLEARMiner.

Nesse cenário, o foco deste trabalho é a mineração de regras temporais multivariadas a partir de dados do comércio internacional. Os resultados automatizados podem auxiliar o economista na tarefa de avaliação e compreensão das informações obtidas. Alguns exemplos de

padrões nesse contexto são: "se aumenta a exportação, o volume de importação também aumenta no mesmo ano", "quando a importação cai e o PIB cai, no mesmo ano a exportação também cai".

1.2 Objetivos do Trabalho

O objetivo principal deste trabalho de mestrado é desenvolver um método de mineração de regras temporais multivariadas para séries temporais econômicas, a ser aplicado e avaliado em dados reais relacionados ao comércio internacional. Visando alcançar esse objetivo e considerando as características dos dados de interesse, as questões de pesquisa a serem respondidas são:

1. Qual o impacto do método de discretização aplicado ao conjunto de séries temporais multivariadas do comércio internacional, considerando as métricas de suporte e confiança, e a interpretabilidade das regras geradas?

2. Qual o impacto da mineração de regras temporais a partir de séries econômicas multivariadas com variáveis e observações faltantes nas regras retornadas e nas métricas de suporte e confiança?

3. Como obter regras confiáveis e coerentes com cenários da economia mundial minerando regras temporais multivariadas dos dados do comércio internacional?

A partir das respostas a essas questões e do desenvolvimento de uma solução, espera-se expandir as contribuições na área de mineração de regras temporais para abranger séries temporais multivariadas heterogêneas e incompletas, propondo e implementando uma solução.

1.3 Contribuições

A principal contribuição deste trabalho é o *extended Temporal RUles Miner* (eTRUMiner), um novo algoritmo capaz de extrair conhecimento de séries temporais multivariadas considerando 2 ou mais variáveis por regra. Uma versão preliminar do algoritmo, denominada TRUMiner, realiza a mineração de regras multivariadas com apenas duas variáveis. O eTRUMiner retorna regras temporais multivariadas até a janela temporal pré-definida pelo usuário, ordenadas pelas métricas de avaliação.

O eTRUMiner suporta séries temporais multivariadas heterogêneas incompletas, com observações e variáveis faltantes. O algoritmo também informa todos os tempos de ocorrência das regras em cada série, possibilitando uma compreensão mais ampla de cada regra obtida. Além disso, o eTRUMiner permite múltiplas discretizações possibilitando a sua aplicação a conjuntos de dados de diversas áreas.

A análise do eTRUMiner sobre o conjunto de dados econômicos do comércio internacional mostrou experimentalmente que a aplicação do algoritmo sobre séries heterogêneas e incompletas é uma alternativa factível para a mineração desse tipo de dado. Com a análise semântica, verificou-se que as regras temporais permitem uma melhor compreensão do período temporal a que se refere, agregando mais informação ao conhecimento obtido.

Dois resultados desse mestrado foram publicados, em Karasawa e Sousa (2022) introduz-se o algoritmo TRUMiner, desenvolvido para minerar regras temporais com 2 variáveis de séries temporais multivariadas heterogêneas e incompletas, enquanto em Karasawa e Sousa (2023) apresenta-se o algoritmo com mais detalhes e novos resultados. A produção científica completa resultante deste trabalho de mestrado, incluindo artigos publicados e códigos do algoritmo implementado, eTRUMiner, estarão disponíveis. O endereço de acesso [4] e *e-mail* de contato [5] podem ser utilizados para maiores informações.

1.4 Organização

Neste capítulo foi introduzido o tema deste trabalho de mestrado, a mineração de regras multivariadas aplicada ao comércio internacional. Apresentou-se a contextualização, os objetivos principais e uma visão geral da abordagem utilizada e contribuições deste trabalho. O restante da dissertação está dividida em 4 capítulos e um apêndice, sumarizados a seguir.

No Capítulo 2 são apresentados os principais conceitos relacionados a mineração de regras temporais. As definições e principais medidas de avaliação de mineração de regras são introduzidas. Em seguida, é abordada a tarefa de mineração de regras temporais, com introdução teórica, seus respectivos trabalhos e estado da arte.

O algoritmo eTRUMiner desenvolvido durante este projeto de mestrado é detalhado no Capítulo 3, com desenvolvimento teórico. Cada etapa do algoritmo é detalhada, discutindo-se abordagens utilizadas e as principais operações realizadas. Questões de implementação, complexidade temporal e espacial da solução são discutidas. No final do capítulo o algoritmo é apresentado.

O Capítulo 4 apresenta o conjunto de dados utilizado, com suas características, e a análise experimental realizada. A avaliação divide-se em três aspectos principais: resultados da aplicação do eTRUMiner sobre o conjunto de dados, capacidade de lidar com a presença de dados faltantes e análise semântica das regras temporais multivariadas mineradas.

Conclui-se este trabalho com o Capítulo 5, destacando-se contribuições e avanços. Os possíveis trabalhos futuros são detalhados nesse capítulo, com a sugestão de algumas abordagens. O Apêndice A apresenta os países e territórios a que se referem as séries do conjunto de dados analisado.

[4] linktree <https://linktr.ee/ekarasawa>
[5] e-mail eligniechk@gmail.com

CAPÍTULO

2

MINERAÇÃO DE REGRAS EM SÉRIES TEMPORAIS

De acordo com Box *et al.* (2015), Morettin e Toloi (2006), uma série temporal pode ser definida como um conjunto de observações temporais que são ordenadas no tempo. No contexto de mineração de dados, as séries temporais frequentemente constituem-se de medidas reais variando em intervalos de tempo regulares (MITSA, 2010), sendo caracterizadas como séries temporais discretas. Analogamente, medidas que variam no tempo de modo contínuo são definidas como séries temporais contínuas.

Além da diferenciação anterior, uma série temporal pode ser composta por observações referentes a uma ou várias variáveis, sendo denominada respectivamente uni ou multivariada. Uma série temporal s discreta multivariada é denotada por

$$s = \{obs_1, obs_2, \ldots, obs_n\}$$

sendo obs_i um vetor δ-dimensional, δ a quantidade de variáveis, $i \in [1, ..., n]$ e n o número de observações da série s. Cada dimensão da i-ésima observação contém o valor obs_i^X de uma variável var_X, cujo domínio pode ser numérico ou categórico, no tempo t_i.

A aplicação do processo de KDD sobre um conjunto de dados S contendo N séries temporais multivariadas pode ser realizada utilizando, por exemplo, a tarefa de mineração de regras temporais. As regras temporais são um caso particular da mineração de regras de associação, contendo informação temporal dos padrões extraídos. O presente capítulo visa introduzir a tarefa de mineração de regras temporais, apresentando conceitos, métodos e trabalhos relacionados.

Nas seções a seguir, serão introduzidas as tarefas de mineração de regras de associação (Seção 2.1), com detalhamento da mineração de regras temporais (Seção 2.2) que é o enfoque deste trabalho. Definições e principais algoritmos são apresentados para melhor compreensão

dessas tarefas. Por fim, são elencados os principais trabalhos relacionados a mineração de regras temporais (Seção 2.3) e apresenta-se as considerações finais (Seção 2.4).

2.1 Regras de Associação

Uma regra de associação é definida por uma relação entre dois elementos em que o antecedente leva ao consequente. A descoberta de regras de associação está inserida no contexto de mineração de padrões, apresentando relevância pelo conhecimento extraído do conjunto de dados e por sua explicabilidade, o que pode facilitar a análise do especialista da área de aplicação.

A mineração de regras de associação (AGRAWAL; IMIELIŃSKI; SWAMI, 1993) busca explicar ou predizer o comportamento dos dados. Formalmente, trata-se de uma implicação do tipo $A \rightarrow C$ com um valor de suporte e confiança. O antecedente da regra (A) e o consequente (C) constituem *itemsets* que são transações pertencentes à lista de itens I, sendo $A \cap C = \emptyset$ e $A, C \neq \varnothing$.

O suporte e a confiança são medidas clássicas usadas para avaliar as regras de associação. Dado o total de transações no conjunto de dados, o suporte (sup) é a quantidade de transações em que o antecedente e o consequente estão presentes, conforme apresentado na Equação 2.1. Se $sup(A \rightarrow C) \geq sup_{min}$, sendo sup_{min} um valor de suporte mínimo definido pelo usuário, diz-se que a regra é frequente.

$$sup(A \rightarrow C) = P(A \cup C) \tag{2.1}$$

A confiança indica o percentual de vezes que a regra ocorre no conjunto de dados quando o antecedente ocorre, calculado pela Equação 2.2. Dado um valor mínimo de confiança fornecido pelo usuário ($conf_{min}$), denomina-se confiável uma regra tal que $conf(A \rightarrow C) \geq conf_{min}$. Quando uma regra de associação é frequente e confiável, ela recebe a denominação de regra forte.

$$conf(A \rightarrow C) = P(C \mid A) \tag{2.2}$$

Existem outras medidas que podem auxiliar na avaliação dos resultados da mineração de regras de associação. O fator de certeza (BERZAL *et al.*, 2002) indica a variação da probabilidade do consequente estar em uma transação quando sabe-se que o antecedente já está. Valores positivos indicam aumento na probabilidade do consequente estar, enquanto negativos indicam redução. Já a informatividade, também conhecida como medida J, fornece a importância de uma regra em termos da informação provida (SMYTH; GOODMAN, 1992).

Avaliar os resultados da mineração de regras de associação pode ser trabalhoso e por vezes necessitar de conhecimento da área de aplicação. Métricas como o suporte e a confiança

podem auxiliar na avaliação e reduzir o trabalho demandado do especialista. Deve-se ter em vista, contudo, que as implicações obtidas indicam correlação e não necessariamente causa entre o antecedente e o consequente e portanto, a validação do conhecimento extraído requer auxílio do especialista.

O algoritmo *Apriori* (AGRAWAL; IMIELIŃSKI; SWAMI, 1993), precursor na extração de regras de associação, é amplamente aplicado e utilizado como base para desenvolvimento de outros algoritmos devido à sua simplicidade. A propriedade *apriori*, na qual se baseia o algoritmo, afirma que todo subconjunto não-vazio de um *itemset* frequente também é frequente. Sua validade é dada pela propriedade anti-monotônica do suporte, na qual o suporte de um *itemset* é sempre menor ou igual ao suporte de seus subconjuntos.

Para reduzir o espaço de busca na geração de *itemsets* frequentes, o algoritmo *Apriori* utiliza a propriedade *apriori* impedindo que a combinação de subsequências infrequentes receba novos itens para formar sequências maiores. A obtenção de regras de associação fortes é realizada utilizando os *itemsets* frequentes encontrados na iteração precedente, gerando todas as regras possíveis. As regras geradas que apresentam confiança abaixo de $conf_{min}$ definido pelo usuário são eliminadas.

O algoritmo *Apriori* desconsidera o fator temporal que pode existir entre os *itemsets* e, embora reduza o processo de geração de regras utilizando a propriedade *apriori*, ainda avalia muitas possibilidades, com múltiplas passagens sobre a base de dados. Novos algoritmos, como o AprioriHybrid (AGRAWAL; SRIKANT, 1994), tentam solucionar o problema das múltiplas passagens.

2.2 Regras Temporais

A mineração de regras temporais, ou regras de associação temporais, é um caso particular da tarefa de mineração de regras de associação, contendo adicionalmente uma característica temporal da regra. Formalmente, uma regra temporal é definida como um par $(A \rightarrow C, \Delta t)$ onde Δt é a característica temporal da regra $A \rightarrow C$ (CHEN; PETROUNIAS, 2000). Um exemplo de conhecimento relacionado a uma regra temporal no contexto de comércio internacional é "quando valores de importação aumentam, os valores de exportação também aumentam no mesmo período". Assim como nas regras de associação, as medidas de suporte e confiança podem avaliar as regras temporais obtidas, devendo serem adaptadas para esse formato.

Para aplicar a tarefa de mineração de regras temporais a conjuntos de séries temporais usualmente é necessária uma etapa de pré-processamento, que pode incluir limpeza, integração e redução dos dados, normalização, discretização e extração de *shapelets*, tratamento de valores faltantes, entre outras operações que visam tornar as tarefas de mineração posteriores mais eficientes (HAN; PEI; KAMBER, 2011). Neste trabalho foram exploradas diferentes técnicas de discretização, que podem ser utilizadas no algoritmo proposto no Capítulo 3.

As características mais detalhadas da mineração de regras temporais incluindo tipos de conjunto de dados utilizado e métricas de avaliação são detalhados a seguir. Na seção seguinte, aborda-se algumas técnicas de discretização aplicáveis que foram utilizadas neste trabalho.

2.2.1 Características

De acordo com Segura-Delgado *et al.* (2020), a mineração de regras temporais pode ser dividida em duas grandes categorias: com o tempo considerado como componente implícito, em que a variável temporal serve para ordenar os dados ou como um fator de relevância, ou com o tempo como componente integral, participando integralmente do processo de mineração. Um exemplo da primeira categoria são regras temporais sem retorno da característica temporal, indicando apenas que o antecedente ocorre antes do consequente. Quando o tempo é um componente integral, ele pode definir a periodicidade ou ser um componente integrante da regra. A proposta deste trabalho é utilizar a informação temporal das séries como característica integral da regra, de forma que duas regras temporais com características temporais distintas são consideradas regras diferentes.

O conjunto de séries temporais aplicado sobre a tarefa pode ser composto por séries univariadas ou multivariadas, contendo múltiplas variáveis. Ao trabalhar com séries temporais multivariadas, pode-se categorizá-las em "homogêneas" ou "heterogêneas" referente a duração de cada variável. Nas séries "homogêneas", todas as variáveis possuem observações contidas no mesmo intervalo temporal. Nas séries "heterogêneas", cada variável da série possui uma duração e portanto, o tempo da observação inicial e final de cada variável não é coincidente.

A distinção entre séries multivariadas "completas" e "incompletas", dá se pela ausência de observações e variáveis ao longo de sua duração. Uma série multivariada "completa" contém observações em todos os intervalos temporais em todas as variáveis. Contudo, séries "incompletas" apresentam valores faltantes como observações pontuais ou ainda em toda a variável, de modo que não há nenhuma observação para aquela variável na série.

A avaliação de regras temporais pode ser realizada utilizando as métricas convencionais de avaliação de regras de associação, o suporte e a confiança. Em trabalhos de mineração de regras temporais que a característica temporal é um componente implícito da regra (HARMS; DEOGUN, 2004; ROMANI *et al.*, 2010), sendo uma janela máxima ou fixa de ocorrência do consequente e que não influencia na diferenciação entre duas regras com o mesmo antecedente e consequente, utiliza-se o suporte e a confiança conforme Equação 2.1 e Equação 2.2 respectivamente. Contudo, essas métricas não avaliam a característica temporal.

2.2.2 Discretização

Na etapa de discretização aplicada sobre as séries temporais, as observações são transformadas em novos valores ou símbolos, de modo a agregar mais contexto, sendo particularmente

útil para tarefas de classificação e mineração de regras. O processo pode ser realizado manualmente, utilizando conceitos externos aos dados ou utilizar características intrínsecas, como as detectadas, por exemplo, por meio de um agrupamento.

Em séries temporais multivariadas, o processo de discretização pode ser aplicado considerando cada uma das variáveis individualmente. Desse modo, dada uma série temporal multivariada s, a discretização da série univariada $s[var_X] = obs_1^X, \ldots, obs_n^X$, em que var_X refere-se à variável da série s, gera uma série temporal discretizada tal que $s'[var_X] = \alpha_{t_1,t_f}^X, \ldots, \alpha_{t_i,t_n}^X$, com L indicando a quantidade de elementos discretizados em $s'[var_X]$. O elemento discretizado compõe-se de um símbolo e o tempo inicial e final das observações que ele abrange. A discretização é representada por

$$s[var_X] = obs_1^X, \ldots, obs_n^X \rightarrow s'[var_X] = \alpha_{t_1,t_f}^X, \ldots, \alpha_{t_i,t_n}^X,$$

com α_{t_i,t_f} referindo-se a um elemento discretizado com tempo inicial e tempo final indicados de forma genérica como t_i e t_f, respectivamente. O elemento discretizado pode representar desde a fração de uma observação até um conjunto de múltiplas observações consecutivas. O tempo inicial (t_i) do elemento discretizado refere-se ao tempo da primeira observação que ele descreve enquanto o tempo final (t_f) é o tempo da última observação considerada.

Durante o processo de discretização, também é possível aplicar uma normalização sobre os dados. As duas principais técnicas de normalização são a *min-max* e a *z-score*. Na normalização *min-max* os valores são reescalados para estarem contidos dentro de um intervalo $[val_{min} - val_{max}]$ pré-definido. A *z-score* normaliza utilizando a média e o desvio padrão do intervalo, mantendo-se um formato mais próximo ao da série original (MITSA, 2010).

2.2.2.1 Comportamental

A discretização denominada "comportamental" neste trabalho consiste em transformar pares de observações consecutivas em padrões de aumento, decréscimo e estabilidade. A finalidade dessa discretização é avaliar a observação em relação ao seu valor anterior, indicando a tendência da série. Para uma série temporal com n observações, são obtidos $L = n - 1$ elementos discretizados.

O número máximo de elementos discretizados distintos gerados pela discretização comportamental é apenas 3, tratando-se de uma discretização com baixo grau de diferenciação entre os valores. O foco dessa discretização é o comportamento geral dos dados na série, evitando a geração exagerada de elementos discretizados distintos.

2.2.2.2 Decis

A categorização de valores de 10% em 10%, denominada "decis", avalia a variação percentual entre observações consecutivas de uma série temporal em grupos de 10%. O percentual é calculado sobre o valor da observação no tempo t_i a partir da diferença entre esta e a observação

no tempo t_{i+1}. A série discretizada pelo método decis possui $L = n - 1$ elementos discretizados, sendo n o número de observações da série original.

A discretização decis fornece um maior detalhamento da série original, indicando o percentual de variação entre observações em nível de decis. Contudo, não há um limite máximo para o número de elementos discretizados distintos a serem gerados pelo método, o que pode dificultar a avaliação da série discretizada. Uma série temporal com n observações pode gerar de 1 até $n - 1$ elementos discretizados distintos, a depender apenas dos valores das observações da série.

2.2.2.3 Quartis

A avaliação de variação em quartis é realizada pela discretização denominada "quartis". Assim como a discretização decis, a quartis indica a variação percentual entre pares de observações consecutivas na série temporal. Cada elemento discretizado refere-se a uma estabilidade, um aumento ou queda em múltiplos de 25% da segunda observação em relação à primeira. A partir de uma série temporal com n observações, a série discretizada por esse método possui $L = n - 1$ elementos discretizados.

O número de elementos discretizados distintos que o método pode gerar depende apenas dos valores das observações da série, assim como na discretização decis. A diferença entre os dois métodos está na quantidade máxima de elementos discretizados distintos gerados para um mesmo intervalo. Entre um aumento de 1 a 100%, por exemplo, a discretização decis pode gerar até dez elementos discretizados distintos enquanto a quartis gera no máximo quatro.

As discretizações comportamental, decis e quartis indicam o perfil de variação entre observações, sendo que os dois últimos métodos fornecem um maior detalhamento da série original, informando respectivamente o decis ou quartis de variação. Esses métodos são de fácil interpretação e facilitam a localização das observações na série temporal referentes ao elemento discretizado.

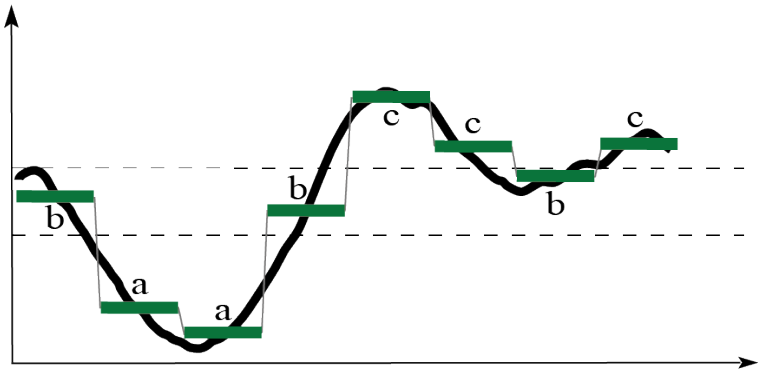
2.2.2.4 SAX

A discretização *Symbolic Aggregate Approximation* (SAX) transforma a série temporal em uma série menor utilizando o *Piecewise Aggregate Approximation* (PAA) como uma técnica intermediária (LIN *et al.*, 2003; LIN *et al.*, 2007). O PAA (KEOGH *et al.*, 2001; KEOGH; PAZZANI, 2000) consiste em redimensionar a série temporal com n observações para uma série de k segmentos de tamanho igual e pré-definido, com cada intervalo apresentando o valor médio de suas observações.

Supondo que as séries temporais normalizadas apresentam uma distribuição Gaussiana, o SAX baseia-se no princípio de equiprobabilidade. Cada série é dividida em trechos de igual

probabilidade e designa-se um símbolo para cada um deles a partir de um alfabeto pré-definido. A geração dos símbolos é feita a respeitar a equiprobabilidade.

Figura 1 – Representação da técnica SAX.



Fonte: Adaptada de Lin *et al.* (2003).

A Figura 1 apresenta a série original em preto. Os trechos em verde foram obtidos após a aplicação do PAA e associação de cada trecho a um símbolo do alfabeto SAX. No exemplo da Figura 1, a série discretizada contém 8 elementos, portanto, $k = 8$. O tamanho do alfabeto utilizado foi igual a três, verificando-se na Figura 1 três símbolos distintos: a, b, c.

Na discretização SAX, o número de elementos discretizados que compõem a série discretizada e a quantidade de símbolos distintos gerados são parâmetros a serem pré-definidos. Portanto, o nível de "semelhança" entre observações e o número de símbolos a serem avaliados pelo processo de mineração pode ser facilmente regulado.

A compreensão da série discretizada pelo método SAX é um pouco mais difícil já que a discretização é influenciada pelos valores mínimo e máximo da série temporal. Contudo, trata-se de uma discretização comumente utilizada (TAN *et al.*, 2015).

2.3 Trabalhos Relacionados

Não foi encontrado nenhum trabalho de mineração de regras temporais aplicada ao comércio internacional. Usualmente, a análise desses dados é realizada através da econometria. Os trabalhos apresentados a seguir referem-se à tarefa de mineração de regras temporais.

Das *et al.* (1998) propuseram um algoritmo para obtenção de regras baseadas em padrões sequenciais intra-série e inter-séries temporais. No trabalho, as séries são discretizadas por meio do "agrupamento de janelas", que consiste na formação de subsequências das séries com o método de janela deslizante seguido de agrupamento. Cada grupo é associado a um símbolo e os símbolos são utilizados para a formação das sequências. A regra é formada por um padrão sequencial e o intervalo temporal é obtido a partir do número de símbolos entre antecedente e consequente. Esse trabalho é um dos precursores na introdução da característica temporal, apresentando uma discretização que evita a necessidade de especialista da área para a tarefa.

O algoritmo MOWCATL, proposto por Harms e Deogun (2004), realiza a mineração de regras temporais baseadas em elementos de interesse fornecidos como parâmetros de entrada, juntamente com a série temporal discretizada e o valor da janela temporal máxima ou janela fixa de tempo. O trabalho é apresentado em mais detalhes na Subseção 2.3.1.

Nam, Lee e Lee (2008) propuseram o algoritmo TARM para a mineração de regras temporais multivariadas sem observações ou variáveis faltantes sobre uma única série temporal. O trabalho possui enfoque na área biológica, visando avaliar as dependências existentes na expressão dos genes. As principais características do algoritmo são discutidas na Subseção 2.3.2.

O CLEARMiner, proposto por Romani *et al.* (2010) para séries agrometeorológicas, extrai regras temporais existentes em uma série temporal multivariada incompleta. As ocorrências das regras temporais podem ser retornadas no formato "extenso" da regra. O CLEARMiner é descrito em mais detalhes na Subseção 2.3.3.

A mineração proposta pelo CMRules (FOURNIER-VIGER *et al.*, 2012) foca apenas na relação sequencial existente entre o antecedente e o consequente da regra temporal. O algoritmo minera regras temporais de bases sequenciais acima dos parâmetros de suporte mínimo sequencial e confiança mínima sequencial. Embora o algoritmo forneça regras temporais mais genéricas, a característica temporal da regra não é definida nem retornada.

O grande volume de regras temporais geradas pelo processo de mineração é tratado em trabalhos mais recentes, com abordagens como a imposição de intervalo mínimo entre antecedente e consequente ou utilização de outros formatos de armazenamento dos dados. Em Wang *et al.* (2018), propõe-se o uso de árvores de *itemsets* frequentes na mineração de regras, visando melhorar a eficiência e a interpretabilidade da tarefa de mineração. O método proposto evita a geração de *itemsets* candidatos e reduz o custo computacional da passagem no conjunto de dados.

Para lidar com o crescimento das séries temporais após a mineração de regras temporais, ou seja, recebimento de novas observações, Wang, Gui e Xu (2022) propõem o IPSTAR, um algoritmo que minera regras temporais compostas por sequências e cuja característica temporal constitui-se o intervalo de acontecimento do consequente. Contudo, durante o processo de discretização, o algoritmo realiza agrupamento das subsequências para determinar os símbolos, que é criticado em Keogh e Lin (2005) por gerar resultados randômicos.

2.3.1 MOWCATL

O MOWCATL não realiza a discretização das séries temporais, focando apenas na tarefa de mineração de regras temporais. Inicialmente são buscados os símbolos de interesse dentro da série multivariada já discretizada fornecida como entrada e selecionados aqueles que possuem suporte mínimo, formando os *itemsets* de antecedentes. *Itemsets* maiores são gerados a partir dos *itemsets* menores desde que mantenham suporte acima do mínimo e a duração esteja dentro da

duração máxima do *itemset*. O processo se repete para os consequentes, com seus respectivos símbolos de interesse.

Os *itemsets* de antecedentes são combinados aos *itemsets* de consequentes, computando-se a diferença temporal entre o início do antecedente e o início do consequente que foram combinados para formar a regra. As regras retornadas são multivariadas limitadas a duas variáveis. O suporte e a confiança das regras são avaliados e o algoritmo retorna as regras com as métricas de avaliação acima das delimitadas pelo usuário. O intervalo temporal exato em que uma regra ocorre nas séries não é retornado, nem as ocorrências das regras nas séries temporais.

2.3.2 TARM

A mineração de regras temporais multivariadas com característica temporal exata de ocorrência é proposta em Nam, Lee e Lee (2008). Aplicada ao estudo da expressão de genes, o trabalho propõe a construção da transação como uma composição de padrões que formam os antecedentes das regras e os consequentes, já vinculada a um intervalo temporal. Dessa forma, evita-se a necessidade de combinação fatorial na geração de regras multivariadas.

O algoritmo *Temporal Association Rule Mining* (TARM) é aplicado sobre uma única série multivariada homogênea e completa, discretizada de modo similar à discretização comportamental. Não há tratamento de observações faltantes ou variáveis desbalanceadas. A avaliação das regras é realizada através do suporte e da confiança mas não é apresentada nenhuma fórmula específica para as regras temporais. As regras retornadas possuem símbolos de duas ou mais variáveis, mas não informam as suas ocorrências nas séries temporais originais.

2.3.3 CLEARMiner

O *CLimatE Association patteRns Miner* (CLEARMiner) foi proposto por Romani *et al.* (2010) para mineração de regras temporais existentes em uma série temporal multivariada incompleta, podendo existir uma janela temporal entre antecedente e consequente da regra. O trabalho foi concebido para dados agrometeorológicos, mas pode ser generalizado para outras aplicações.

O algoritmo divide-se em duas etapas principais: discretização e extração de regras. A entrada constitui-se das séries temporais de interesse, os limites mínimos utilizados para discretização e a janela de análise das regras. Pode-se fornecer também o suporte e a confiança mínimos.

A rotina de discretização é a principal responsável por evitar que o número de regras geradas seja extenso. Para tal, seleciona-se apenas os *itemsets* que estão acima de todos os limites mínimos para discretização e possuem suporte acima do mínimo definido pelo usuário. A extração das regras utiliza os *itemsets* gerados e avalia a confiança de todas as regras possíveis dentro da janela definida pelo usuário, retornando aquelas acima do limite mínimo.

As regras temporais são geradas no formato "curto", contendo antecedente e consequente, e "extenso", em que são fornecidas as observações iniciais, intermediárias e finais e o tempo inicial e final do antecedente e do consequente da regra nas séries de ocorrência. Um exemplo é a regra "Chuva [0, 45, 0] (01/10/2010, 01/12/2010) → Temperatura [27, 22, 25] (01/10/2010, 01/12/2010)" apresentada em Romani *et al.* (2013), indicando que um pico de chuva é seguido por uma queda de temperatura, comportamento verificado entre outubro e dezembro de 2010. Assim, os *itemsets* pertencentes às regras são diretamente relacionados às respectivas séries temporais.

2.4 Considerações Finais

A mineração de regras temporais é a tarefa de mineração de séries temporais aplicada neste trabalho de mestrado. Uma regra temporal pode ser visualizada como uma extensão de uma regra de associação, e por tanto, as regras de associação são explicadas na Seção 2.1 com menção ao algoritmo *Apriori*, precursor da área.

As principais abordagens para minerar regras temporais variam quanto ao volume e formato dos dados de entrada, parâmetros que serão impostos pelo algoritmo e formato das regras que será retornado. Trabalhos correlatos a este projeto de mestrado foram descritos na Seção 2.3. O Quadro 1 apresenta uma comparação dos trabalhos existentes mais próximos à pesquisa realizada nesse projeto de mestrado. Como destacado na Seção 2.3, não foi encontrado nenhum trabalho com essa abordagem para a análise do comércio internacional. Portanto, os trabalhos no Quadro 1 assemelham-se quanto à solução proposta.

Quadro 1 – Comparação do eTRUMiner e trabalhos relacionados.

Trabalho	Dado de Entrada	Tipo de Série	Volume de Séries	Regra Multivariada	Característica Temporal	Localização da Regra
MOWCATL, 2004	SÉRIE DISCRETIZADA	COMPLETA	SÉRIE ÚNICA	2	NÃO	NÃO
TARM, 2008	SÉRIE TEMPORAL	COMPLETA	SÉRIE ÚNICA	2+	SIM	NÃO
CLEARMiner, 2010	SÉRIE TEMPORAL	INCOMPLETA	SÉRIE ÚNICA	2	NÃO	SIM
eTRUMiner	SÉRIE TEMPORAL	INCOMPLETA	MÚLTIPLAS SÉRIES	2+	SIM	SIM

Fonte: Elaborada pela autora.

O MOWCATL (HARMS; DEOGUN, 2004) extrai regras temporais de uma única série multivariada já discretizada. A série deve ser completa, também devendo ser informado símbolos de interesse para compor as regras. A característica temporal não é um componente integral da

regra, sendo que as regras retornadas não possuem informação temporal e compõem-se de até 2 variáveis distintas.

O método TARM (NAM; LEE; LEE, 2008) propõe uma nova construção de transação a partir da série temporal, facilitando a mineração de regras com 3 ou mais símbolos de variáveis distintas. A entrada do algoritmo é uma única série temporal completa, retornando regras com a característica temporal mas sem localização das ocorrências na série.

O algoritmo CLEARMiner (ROMANI *et al.*, 2010) inclui a discretização em sua rotina, permitindo a mineração de séries temporais incompletas. A extração de regras é realizada sobre uma única série multivariada, e as regras retornadas compõem-se de até duas variáveis distintas. O CLEARMiner é o único método que informa a localização de ocorrências das regras nas séries através do formato "extenso", mas a característica temporal da regra não é informada.

Neste trabalho de mestrado desenvolveu-se o algoritmo eTRUMiner capaz de minerar de regras temporais multivariadas, contendo 2 ou mais variáveis, a partir de um conjunto de múltiplas séries multivariadas incompletas relacionadas ao comércio internacional. O algoritmo, que integra a etapa de discretização das séries, informa a ocorrência de cada regra nas séries temporais originais. Nesse contexto, não foi localizado nenhum trabalho similar. No próximo capítulo será detalhado a solução implementada, o algoritmo eTRUMiner. Conceitos, algoritmo e complexidade são abordados para melhor compreensão.

CAPÍTULO

3

ETRUMINER

Os dados do comércio internacional tratados neste trabalho são séries temporais univariadas que compõem um conjunto de séries temporais multivariadas heterogêneas e incompletas, ou seja, com valores em escalas distintas e, principalmente, com observações e variáveis faltantes. Devido ao perfil heterogêneo do conjunto de séries (descrito em detalhes na Seção 4.1), e às limitações dos métodos propostos na literatura para mineração de regras temporais (conforme discutido na Seção 2.3), foi desenvolvido neste trabalho o algoritmo eTRUMiner, capaz de minerar regras temporais multivariadas a partir de séries temporais multivariadas heterogêneas e incompletas, sem a necessidade de tratamentos prévios.

3.1 Conceitos

Dado um conjunto S de séries temporais multivariadas

$$S = s_1, \ldots, s_N$$

aplica-se o processo de discretização e obtém-se o conjunto de séries discretizadas

$$S' = s'_1, \ldots, s'_N$$

cada série s_i contendo até δ variáveis, conforme detalhado nas definições do Capítulo 2.

No contexto de mineração de regras temporais, define-se para este trabalho uma transação como sendo

$$([var_X, \alpha_i^X], \ldots, [var_Y, \alpha_j^Y], \ldots), \Delta t,$$

em que cada padrão (var_X, α_i^X) constitui-se de uma variável (var_X) e um elemento discretizado (α_i^X). A transação constitui-se de dois conjuntos de padrões que podem possuir tempos iniciais distintos, o antecedente, aquele que possui o menor tempo inicial, e o consequente com tempo inicial maior ou igual ao tempo inicial do antecedente. Além disso, cada variável pertence a apenas um único padrão da transação, no máximo.

Para reduzir as possíveis regras geradas, cada padrão dentro do mesmo conjunto da transação possui exatamente o mesmo tempo inicial, embora não há necessidade dos tempos finais coincidirem. O fator temporal da transação (Δt) indica a diferença dos tempos iniciais entre antecedente e consequente da transação. Assim como em Romani *et al.* (2010), delimita-se uma janela temporal w que indica a diferença máxima entre antecedente e consequente da transação, tentando gerar apenas transações de relacionamentos causais.

Em um conjunto de séries multivariadas homogêneo e completo, o número de transações é dado por

$$T = \frac{N}{2}.(w+1)(2.L-w)$$

considerando que todas as séries discretizadas possuem L observações discretizadas cada. Contudo, o número de observações discretizadas pode variar de acordo com o método de discretização utilizado.

As regras temporais multivariadas são obtidas a partir do conjunto de transações geradas. Uma regra temporal multivariada genérica é dada por

$$([var_X, \alpha_i^X], \dots) \Rightarrow ([var_Y, \alpha_j^Y], \dots), \Delta t$$

em que $([var_X, \alpha_i^X], \dots)$ compõem o antecedente da regra e o consequente é $([var_Y, \alpha_j^Y], \dots)$. A característica temporal da regra (Δt) é o fator temporal da transação que a gerou, e indica a diferença de tempo do início dos padrões no antecedente e o início dos padrões no consequente.

Denomina-se como uma regra "curta" a regra temporal contendo o antecedente, o consequente e a característica temporal, conforme definido na Seção 2.2. Uma regra "extensa" compõe-se da regra curta (antecedente, consequente e característica temporal) e todas as suas ocorrências nas séries temporais. Para localização de cada ocorrência da regra é necessário o índice da série e o tempo da primeira observação que refere-se ao início da regra temporal.

A regra temporal é formada por padrões de uma única transação, contendo pelo menos um padrão do antecedente da transação no antecedente da regra e um do consequente da transação no consequente da regra. Como cada variável das séries discretizadas compõe apenas um padrão, a regra temporal é composta de até δ variáveis, contendo de 2 a δ padrões.

Para avaliar as regras temporais obtidas, estende-se as métricas de avaliação tradicionais da mineração de regras de associação, o suporte e a confiança. O suporte indica a frequência da regra no conjunto de transações enquanto a precisão da regra é medida pela confiança. A Equação 3.1 apresenta o suporte temporal multivariado usado neste trabalho, com base em (ROMANI *et al.*, 2010).

$$sup = 100.\frac{freq([var_X, \alpha_i^X], \cdots \Rightarrow [var_Y, \alpha_j^Y], \dots, \Delta t)}{T} \tag{3.1}$$

O suporte de uma regra de associação varia de 0 a 100, sendo uma regra com suporte igual a 100 quando é verificada em todas as transações do conjunto. Contudo, na mineração

de regras temporais, cada transação possui um fator temporal que é utilizado na geração de regras. Portanto, nenhuma regra temporal pode estar presente em todas as transações do conjunto quando a janela temporal for maior que zero ($w > 0$). A equação abaixo indica o número de transações com um fator temporal $\Delta t = k$

$$T_{\Delta t=k} = N.(L-k)$$

A quantidade percentual de transações com fator temporal $\Delta t = k$ no total de transações é dada pela Equação 3.2.

$$T_{\Delta t=k}(\%) = 100.\frac{2.(L-k)}{(w+1)(2.L-w)} \tag{3.2}$$

Por exemplo, para séries com 23 elementos discretizados e uma janela temporal $w = 5$, o fator temporal $\Delta t = 0$ está presente em $18,7\%$ das transações, de modo que o suporte máximo que uma regra com característica temporal $\Delta t = 0$ pode obter é $18,7$. Quanto maior o tamanho da janela temporal avaliada, menor será o suporte máximo possível de ser atingido.

A confiança de uma regra temporal, que indica a sua chance de acontecimento, é dada pela frequência da regra dividida pela frequência de todas as transações que geram regras temporais com o mesmo antecedente e a mesma característica temporal. Baseado em (ROMANI *et al.*, 2010), a Equação 3.3 calcula a confiança em regras temporais multivariadas.

$$conf = 100.\frac{freq([var_X, \alpha_i^X], \cdots \Rightarrow [var_Y, \alpha_j^Y], \ldots, \Delta t)}{freq([var_X, \alpha_i^X], \ldots, \Delta t)} \tag{3.3}$$

Como o fator temporal afeta a avaliação do suporte, define-se como uma regra "genérica" o agrupamento de regras temporais no antecedente e consequente. A regra genérica assemelha-se à uma regra de associação, compondo-se de antecedente e consequente, mas é obtida a partir das regras temporais, sendo que o antecedente ocorre apenas junto ou antes que o consequente. A característica temporal da regra temporal que tiver o maior suporte e confiança dentre o grupo de regras que compõem a regra genérica é denominado intervalo temporal principal.

O suporte de uma regra genérica pode ser calculado a partir dos suportes das regras temporais. Dado que trata-se de um agrupamento das regras e o total de transações permanece o mesmo, o suporte de uma regra genérica é simplesmente a soma dos suportes. Para a confiança contudo, é necessário determinar o número de transações com o antecedente da regra genérica desconsiderando o fator temporal. O total de transações multiplicado pela somatória dos suportes das regras temporais dividido pelo número de transações do antecedente informa a confiança da regra genérica.

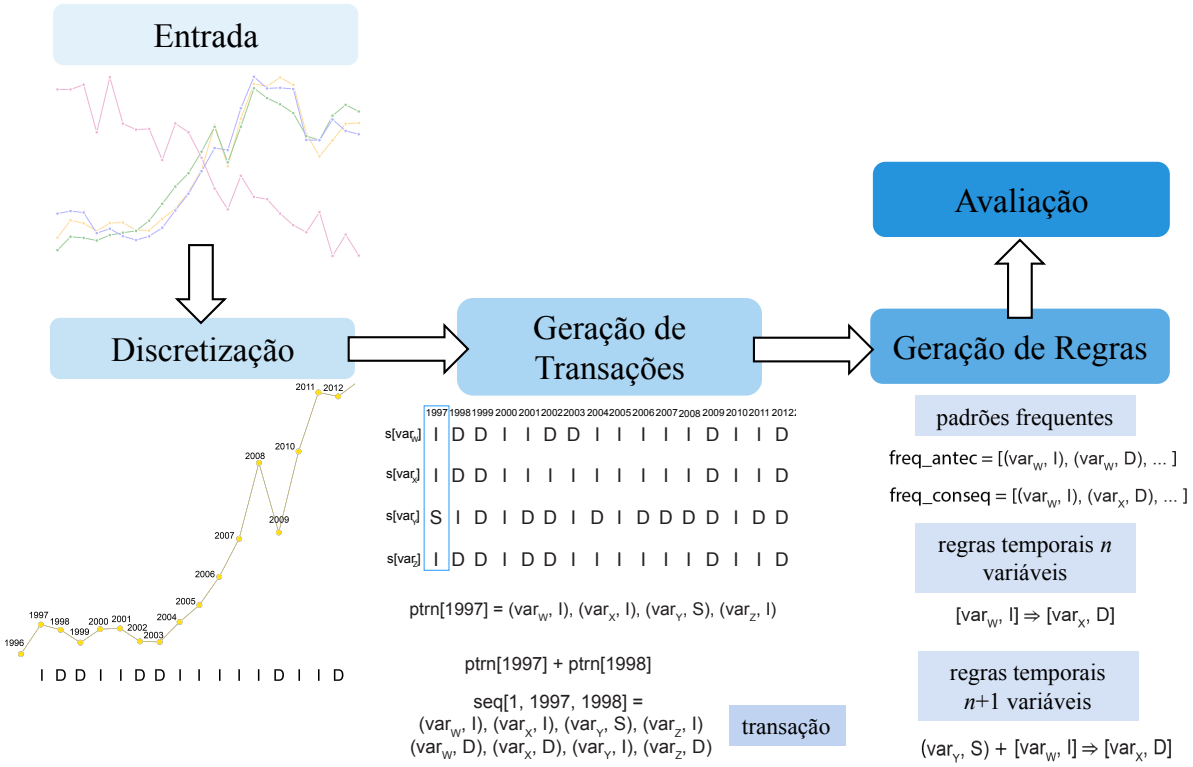
O conjunto de dados de entrada do algoritmo (S) constitui-se de séries temporais univariadas representando cada variável da série multivariada. Cada variável pode ter duração distinta e o eTRUMiner é capaz de lidar com a presença de observações e variáveis faltantes. Para que

as variáveis de uma série multivariada sejam compreendidas como referentes à mesma série, é necessário apenas que o identificador de cada variável da série seja idêntico, independentemente da ordem de entrada entre as séries de uma variável.

Além das séries univariadas, são parâmetros de entrada: o método de discretização (*disc*), a janela temporal (w) para limitar a característica temporal máxima da regra, o suporte mínimo (sup_{min}) e a confiança mínima ($conf_{min}$). A janela temporal é um parâmetro de usuário que limita a diferença de tempo inicial entre os antecedentes e consequentes da regra. Esse parâmetro delimita o intervalo temporal máximo em que ainda há sentido semântico na regra, reduzindo o trabalho computacional de combinação entre todos os padrões possíveis para até o limite semântico de tempo.

Para a obtenção de melhores resultados na mineração, é necessário levar em consideração as características específicas do conjunto de dados analisado. Por exemplo, a discretização aplicada sobre as séries temporais deve possuir sentido semântico, a janela temporal deve levar em consideração o período semântico e os valores de corte nas métricas de avaliação (suporte e confiança mínimos) devem ser delimitados através dos menores suporte e confiança das regras temporais que ainda possuem sentido semântico.

Figura 2 – Exemplo ilustrativo do algoritmo eTRUMiner.



Fonte: Elaborada pela autora.

O algoritmo eTRUMiner, ilustrado na Figura 2 pode ser dividido em quatro etapas principais, detalhadas nas seções a seguir: a discretização, a geração de transações, a geração

de regras e a avaliação. O conjunto de entrada constitui-se de séries temporais multivariadas, representadas na Figura 2 como cores distintas, podendo conter durações distintas, valores e variáveis faltantes.

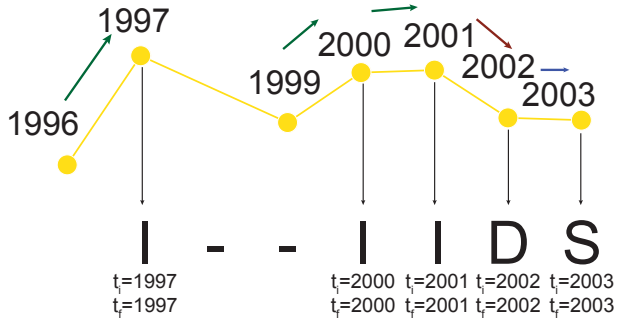
O processo de discretização consiste em transformar as séries temporais em séries com elementos discretizados a partir do método de discretização determinado pelo usuário. Na etapa de geração de transações, os elementos discretizados são associados para formarem transações que são posteriormente utilizadas para obter regras temporais multivariadas compostas dos padrões frequentes na fase de geração de regras. Por fim, as regras temporais são avaliadas na última fase por meio de métricas de avaliação pré-definidas, para retorno de regras relevantes. As operações realizadas em cada etapa são detalhadas nas seções seguintes e sintetizadas no Algoritmo 1.

3.2 Discretização

Para a tarefa de mineração de regras temporais, a escolha do método de discretização deve considerar a área de aplicação do conjunto de dados analisado. O algoritmo eTRUMiner foi desenvolvido para um conjunto de dados econômicos, e portanto, inclui discretizações adequadas para essa aplicação por serem de fácil interpretação dos resultados e/ou comumente utilizada: comportamental, decis, quartis e SAX.

Cada elemento discretizado no algoritmo possui um símbolo que representa seu comportamento, o tempo inicial (t_i) e o final (t_f) da observação ou conjunto de observações referentes, apresentados como subíndices na linha 8 do Algoritmo 1. Ao armazenar o tempo inicial e final, o elemento pode corresponder de parte de uma observação até múltiplas observações consecutivas, sem a necessidade de duração constante entre os elementos discretizados.

Figura 3 – Discretização exemplificada para a presença de observações faltantes.



Fonte: Elaborada pela autora.

O tratamento de observações e variáveis faltantes pelo eTRUMiner é possibilitado pela etapa de discretização. O armazenamento do tempo inicial e final indica com exatidão quais observações o elemento discretizado representa, de forma que os elementos são comparados pelos seus tempos e não pela posição na série discretizada. A Figura 3 apresenta um exemplo do

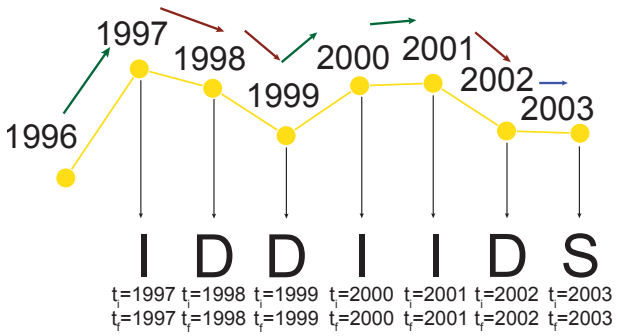
processo de discretização com a observação de 1998 faltante. Na etapa de geração de transações, os elementos discretizados da a série discretizada resultante serão associados utilizando-se o tempo inicial (t_i) e final (t_f), permitindo minerar séries com observações faltantes. Dessa forma, o eTRUMiner simplifica o pré-processamento necessário, permitindo o tratamento de observações e variáveis faltantes sem a necessidade de imputação, além de séries temporais com durações diferentes.

3.2.1 Comportamental

A discretização comportamental mensura a variação entre observações consecutivas (Subsubseção 2.2.2.1). Essa variação é medida pela diferença dos valores absolutos e classificada entre aumento, queda e estabilidade. Iniciando pela segunda observação, para cada par, calcula-se a diferença entre a observação atual e a anterior. Em caso de aumento, o símbolo é "I" (do inglês "increase"), mas se há um decréscimo entre observações consecutivas, o símbolo retornado é "D" (do inglês "decrease"). A estabilidade entre duas observações é representada pelo símbolo "S" (do inglês "stability") e o limite máximo para ser considerado como estabilidade é a diferença máxima de $0,1\%$ entre as observações.

O objetivo da discretização comportamental é representar o comportamento entre duas observações e não a própria observação ou conjunto de observações específico. Dessa forma, o tempo inicial do elemento discretizado indica a observação na série temporal que resultou desse comportamento. Para evitar a compreensão de que o elemento discretizado refere-se desde a observação do tempo inicial até a observação do tempo final, em métodos que representam o comportamento entre observações, fixou-se o tempo final (t_f) igual ao tempo inicial (t_i).

Figura 4 – Discretização exemplificada pelo método comportamental aplicado sobre a série de importação do Brasil.



Fonte: Elaborada pela autora.

A Figura 4 ilustra o procedimento de discretização pelo método comportamental sobre a série de importações do Brasil de 1996 a 2003. Para a discretização comportamental, é necessário pelo menos duas observações consecutivas para gerar um elemento discretizado. O primeiro elemento discretizado apresenta tempo inicial (t_i) e final (t_f) em 1997, indicando que em 1997 houve um aumento nas importações em relação ao valor do ano anterior. O

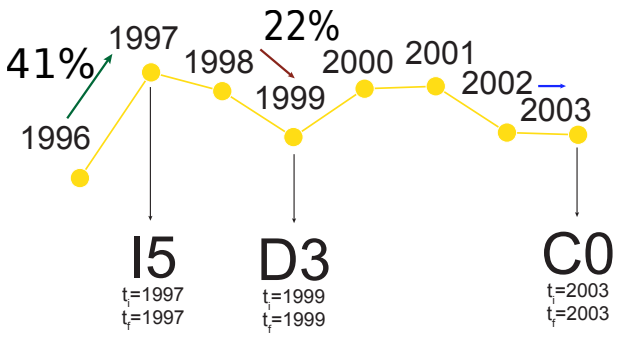
último elemento discretizado refere-se a 2003, o ano da última observação desse exemplo. Na ocorrência de observações faltantes, as discretizações utilizadas geram elementos discretizados até a última observação presente e após a presença de novas observações na série temporal, como exemplificado na Figura 3.

A discretização comportamental realiza uma forte sumarização do comportamento entre observações das séries analisadas, reduzindo em apenas 3 tipos de símbolos que indicam crescimento, decréscimo ou estabilidade. Esse perfil conciso resulta em um menor número de regras que é benéfico, pois demanda menor trabalho para a avaliação do especialista, mas implica uma maior generalização de comportamentos, perdendo características mais específicas.

3.2.2 Decis

A discretização em variações de 10% entre observações consecutivas é realizada pela discretização decis, conforme descrito na Subsubseção 2.2.2.2. Essa variação pode ser sobre a diferença percentual dos valores absolutos ou dos valores normalizados da série univariada. Uma possibilidade de normalização é a *z-score*, que mantém a proporção de variação entre as observações, já contemplada no eTRUMiner, que pode ser utilizada em conjunto com a discretização decis.

Figura 5 – Discretização exemplificada pelo método decis aplicado sobre a série de importação do Brasil.



Fonte: Elaborada pela autora.

O processo é semelhante ao realizado na discretização comportamental, sendo necessário duas observações consecutivas para gerar um elemento discretizado. Inicia-se pela segunda observação e calcula-se para cada par a diferença percentual entre a observação atual e a anterior. Ao contrário da discretização comportamental, a decis não possui um número máximo de elementos discretizados distintos.

Em caso de aumento, o símbolo é "I" e o número do decis que representa o percentual de aumento. A Figura 5 ilustra a discretização decis, com o primeiro elemento discretizado referindo-se a um aumento de 5 decis em relação à observação anterior. O decréscimo apresenta a mesma lógica, com o símbolo "D" concatenado com número do decis de decréscimo. Na Figura 5 o decréscimo de 22% é discretizado em um elemento D3. A repetição idêntica de

valores em duas observações consecutivas é discretizada pelo elemento “C0” (com C do inglês “constant”).

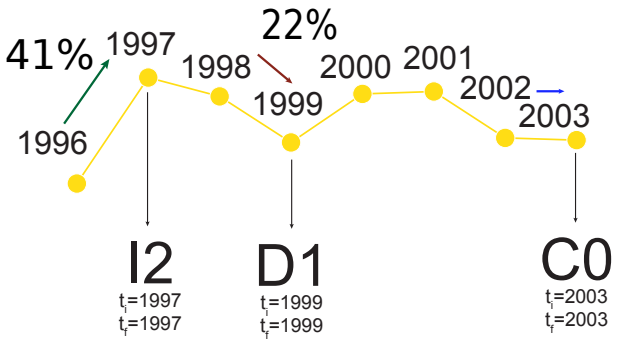
A discretização decis informa a variação percentual entre observações das séries analisadas com detalhe a nível de 10%, que permite melhor compreensão e discernimento entre comportamentos existentes nas séries. Contudo, esse detalhamento gera um grande número de regras, o que pode dificultar a avaliação das regras pelo especialista.

3.2.3 Quartis

A discretização quartis é similar à discretização decis, pois avalia a variação percentual entre observações consecutivas de forma mais detalhada. Porém, as variações são classificadas em quartis, percentuais de 25% (Subsubseção 2.2.2.3). A normalização pode ser utilizada em conjunto com a discretização quartis conforme tratamento dos dados desejado

O cálculo da diferença entre a observação atual e a anterior é realizado a partir da segunda observação, avaliando-se o percentual de variação em relação à observação anterior. Na discretização quartis também não há número máximo de elementos distintos, contudo, para uma mesma faixa de variação entre os valores, a quartis gera até 4 elementos distintos a cada 10 elementos gerados pela discretização decis.

Figura 6 – Discretização exemplificada pelo método quartis aplicado sobre a série de importação do Brasil.



Fonte: Elaborada pela autora.

Em caso de aumento, o símbolo “I” é concatenado com o número do quartis que representa o percentual de aumento. Para a mesma série apresentada nas discretizações anteriores, a Figura 6 ilustra os elementos discretizados. O aumento de 41% é representado por um elemento de I2 (entre 25% e 50% de aumento), o decréscimo de 22% está contido no primeiro quartil e é representado por D1 e a estabilidade entre 2002 e 2003 retorna um elemento discretizado C0.

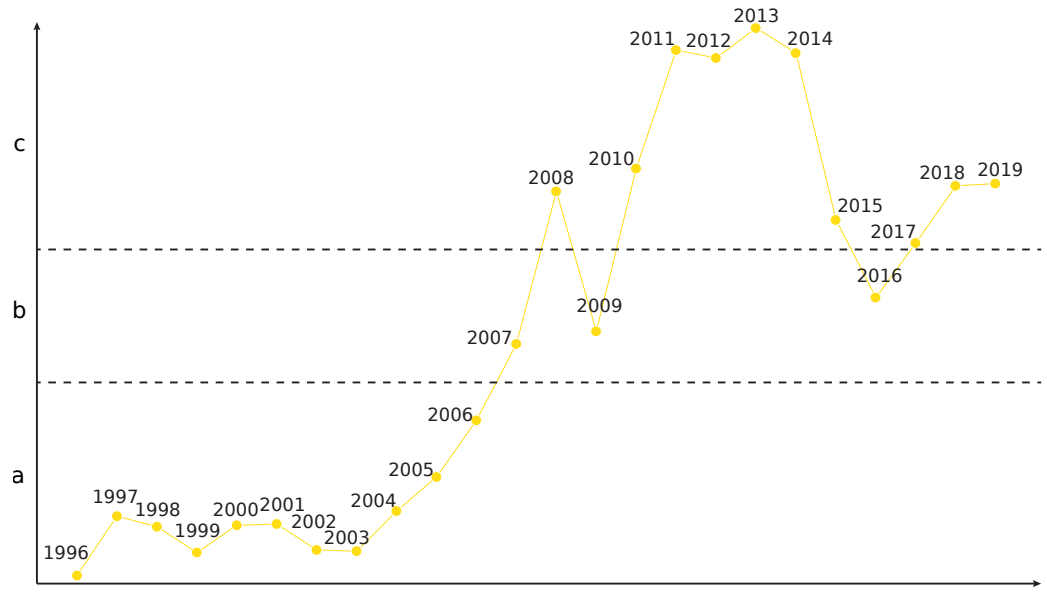
A discretização quartis informa o variação percentual entre observações das séries analisadas com detalhe a nível de quartis. O número de regras geradas em relação à discretização decis é menor, porém, frequentemente maior que a discretização comportamental. Um grande volume de regras pode dificultar a avaliação do especialista, mas permite melhor compreensão

e discernimento entre comportamentos existentes nas séries. Portanto, é importante avaliar o melhor método de discretização a ser utilizado de acordo com os propósitos da análise.

3.2.4 SAX

Para discretização no método SAX, o usuário deve indicar o número final de elementos discretizados desejado (tamanho da série discretizada) e o número de letras a ser utilizado para a discretização, entre um mínimo de 2 e máximo de 26, que é total de elementos discretizados distintos. O tempo inicial e final dos elementos discretizados dependem do parâmetro fornecido pelo usuário que indica a quantidade total desejada de elementos discretizados na série discretizada. A aplicação do SAX com o parâmetro de 3 elementos discretizados distintos, por exemplo, gera os símbolos “a”, “b” e “c” conforme ilustrado na Figura 7.

Figura 7 – Discretização exemplificada pelo método SAX aplicado sobre a série de importação do Brasil.



Fonte: Elaborada pela autora.

A normalização *z-score* é aplicada sobre os dados devido aos requisitos do método de discretização, conforme Subsubseção 2.2.2.4. Em seguida, o método *Piecewise Aggregate Approximation* (PAA) é aplicado sobre a série normalizada para obter o número de elementos desejado na série discretizada. Por fim encontra-se a letra correspondente para cada observação ou conjunto de observações. Letras mais próximas ao início do alfabeto indicam valores menores na distribuição de valores da série, enquanto os valores maiores são representados por letras mais no final do alfabeto.

Como o número de regras distintas está diretamente relacionado ao número de elementos discretizados distintos que o método de discretização pode gerar, ao utilizar a discretização SAX, o usuário pode regular o volume final de regras. Contudo, a interpretação dos elementos de uma série discretizada pelo método SAX é vinculada fortemente à distribuição de valores da série, o que pode dificultar a compreensão das regras.

3.3 Geração de Transações

O objetivo da geração de transações é armazenar para cada intervalo temporal Δt, os padrões que poderão formar os antecedentes e consequentes das regras, assim como a série e os tempos iniciais que constituem cada transação. Dessa forma, a etapa de geração de regras pode ser realizada para cada Δt e seus antecedentes e consequentes específicos.

A definição de transações utilizada pelo eTRUMiner (Seção 3.1) possibilita a mineração de regras multivariadas a partir da transação, pois a regra contém no máximo todos os seus elementos discretizados. Além disso, com o armazenamento dos índices das transações de ocorrência de cada padrão, facilita-se a geração de regras, uma vez que para o cálculo do suporte é necessário apenas quantificar a intersecção de ocorrências de cada padrão que compõe a regra.

Um padrão consiste num elemento discretizado e sua respectiva variável. Após a geração de todas as séries discretizadas na etapa de discretização, percorre-se cada série ordenando os padrões por tempo inicial (t_i) do elemento discretizado (linhas 7 a 11 do Algoritmo 1). Padrões com o mesmo tempo inicial são armazenados na mesma lista, sendo o tamanho máximo da lista delimitado pelo número de variáveis da série.

Figura 8 – Exemplo de padrões armazenados do Brasil ordenados por tempo inicial.

```
ptrn[1997] = (imp, I), (exp, I), (ECI, S), (PIB, I)
ptrn[1998] = (imp, D), (exp, D), (ECI, I), (PIB, D)
ptrn[1999] = (imp, D), (exp, D), (ECI, D), (PIB, D)
ptrn[2000] = (imp, I), (exp, I), (ECI, I), (PIB, I)
ptrn[2001] = (imp, I), (exp, I), (ECI, D), (PIB, D)
```

Fonte: Elaborada pela autora.

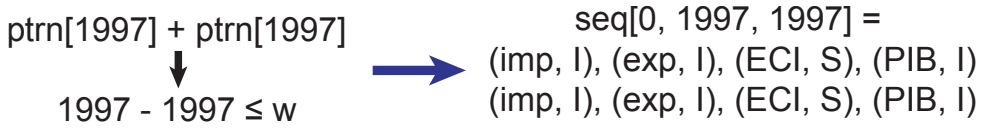
A Figura 8 ilustra os padrões armazenados para a série multivariada do Brasil, com "imp" referindo-se ao valor de importação, "exp" referindo-se ao valor de exportação, "ECI" sendo o Índice de Complexidade Econômica e "PIB" referindo-se ao Produto Interno Bruto. Os elementos discretizados da Figura 8 foram gerados utilizando a discretização comportamental.

Para todos os tempos iniciais contidos em *ptrn* de cada série, suas listas de padrões são concatenadas, dentro do mesmo tempo inicial (t_i) e entre diferentes t_i, desde que a diferença entre os tempos iniciais concatenados não ultrapasse o tamanho da janela temporal definido (w). Esse processo gera as transações que serão utilizadas na geração de regras. Armazena-se a diferença entre os tempos iniciais, denominado fator temporal Δt, e ambos os tempos iniciais, sendo o elemento composto por um par contendo as listas dos padrões concatenados.

O processo de geração de transações a partir de listas de padrões é ilustrado na Figura 9. A transação exemplificada é gerada a partir da lista de padrões com tempo inicial 1997 para formar o conjunto antecedente e também para o conjunto consequente, formando uma transação

com fator temporal 0. A geração de uma transação ocorre apenas se o intervalo temporal entre padrões concatenados for igual ou inferior à janela temporal (Algoritmo 1 linhas 12 - 20).

Figura 9 – Exemplificação do processo de geração de transação realizada sobre série do Brasil.



Fonte: Elaborada pela autora.

No exemplo da Figura 9, a transação pertence ao dicionário *seq* e é caracterizada pelo fator temporal 0, 1997 como tempo inicial do conjunto antecedente e 1997 como tempo inicial do conjunto consequente. Os conjuntos antecedente e consequente são armazenados separadamente, para diferenciação dos conjuntos.

Todos os padrões que compõem os conjuntos antecedentes e consequentes gerados são armazenados unicamente para cada fator temporal da transação (Algoritmo 1 linhas 21 a 32). A Figura 10 ilustra essa etapa, com cada padrão de cada fator temporal armazenando os índices das transações que ele compõe o conjunto antecedente (em uma primeira lista) e o consequente (segunda lista).

Na Figura 10, os padrões da transação $seq[0, 1997, 1997]$ são armazenados em um dicionário do fator temporal 0. As informações da transação (índice da série temporal que se refere, fator temporal, tempo inicial do conjunto antecedente e tempo inicial do conjunto consequente) são armazenadas em uma lista denominada *itemset_info*. O índice de *itemset_info* é inserido no dicionário de padrão, indicando que aquele padrão ocorre na transação de índice *itemset_info*.

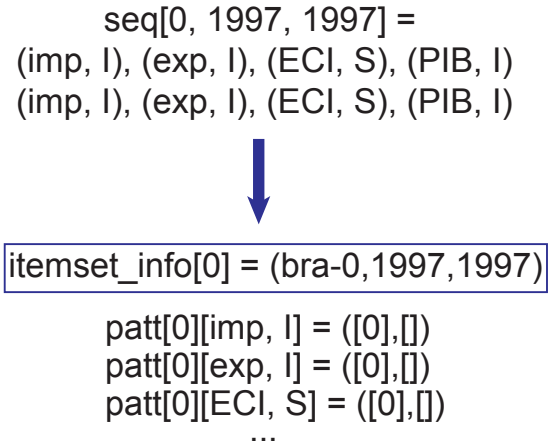
Por exemplo, o padrão $[imp, I]$ pertence ao conjunto antecedente e consequente da transação $seq[0, 1997, 1997]$ que possui índice 0 em *itemset_info*. No dicionário de padrões, representado por *patt*, o padrão $[imp, I]$ terá o índice 0 inserido na primeira lista, que armazena as ocorrências do padrão em transações compondo o conjunto consequente. Cada transação gerada tem seus padrões armazenados em *patt*.

Cada padrão armazena os índices das transações que compõe para facilitar a localização da intersecção dos padrões na formação de regras. Esse método de armazenamento baseia-se na abordagem de representação vertical utilizada por Zaki (2001) para otimizar a tarefa de mineração de sequências. Além disso, cada padrão pode pertencer tanto aos antecedentes quanto aos consequentes que formarão as regras. O Algoritmo 1 linhas 6 - 34 sumariza os processos desta etapa, descritos e ilustrados nesta subseção.

3.4 Geração de Regras

A etapa de geração de regras é realizada após o armazenamento das transações, individualmente para cada intervalo temporal Δt dentro da janela temporal (w) determinada pelo usuário.

Figura 10 – Ilustração do processo de armazenamento de padrões a partir de transações obtidas da série do Brasil.



Fonte: Elaborada pela autora.

Para cada padrão contido nas transações, o número mínimo de ocorrências ($ocorr_{min}$) para ser utilizado na formação de regras é dado por

$$ocorr_{min} = sup_{min} * T \tag{3.4}$$

sendo sup_{min} o suporte mínimo que é parâmetro do usuário, e T o total de transações do conjunto de dados, que pode ser obtido através tamanho do vetor que contém as informações de cada transação.

Para cada padrão avalia-se o número de ocorrências como antecedente de uma transação, armazenado no primeiro item do padrão contido no dicionário de padrões *patt*. Se esse valor for acima do mínimo, adiciona-se o padrão na sequência de antecedentes frequentes que serão utilizados na formação de regras. O mesmo ocorre para formar a sequência de consequentes frequentes (Algoritmo 1 linhas 36 - 43).

Cada padrão dos antecedentes frequentes é unido a cada padrão dos consequentes frequentes desde que a variável do padrão antecedente seja distinta da variável do consequente. O pré-requisito de variáveis distintas foi definido para essa pesquisa e pode ser removido conforme o objetivo de aplicação do eTRUMiner. Verifica-se a intersecção das ocorrências nas séries discretizadas do antecedente com as ocorrências do consequente e se o número de ocorrências na intersecção for maior que o mínimo conforme Equação 3.4, adiciona-se a regra ao conjunto de regras finais (*Srules*) e regras a serem ampliadas (*rules*) (Algoritmo 1 linhas 44 - 53).

O eTRUMiner gera regras com o máximo de variáveis distintas possíveis. Dessa forma, se o conjunto de dados possuir mais que duas variáveis, a mineração de regras continua até atingir o limite de variáveis ou não ser possível gerar mais nenhuma regra distinta. Para cada regra obtida com 2 variáveis, tenta-se inserir um padrão antecedente, verificando se a variável do antecedente a ser inserido é distinta das demais variáveis já presentes na regra e se o número de

ocorrências da intersecção também é acima do mínimo determinado pela Equação 3.4. Em caso de sucesso, a regra é adicionada ao conjunto de regras finais e à lista de regras geradas com um item adicionado.

O processo também se repete para a adição de padrão consequente, sendo que a ordem entre padrões no antecedente e no consequente é definida pela ordem das variáveis informadas na entrada de dados, para evitar a geração de regras idênticas porém com ordem distinta dentro do antecedente ou consequente. Além disso, se uma regra gerada for idêntica a uma existente, a mesma não será adicionada em duplicidade. Enquanto o conjunto de regras geradas na última adição de padrões não for vazio e o número de variáveis nas regras geradas for menor que total de variáveis, continua-se esse processo (Algoritmo 1 linhas 54 - 78).

Após a geração de todas as regras possíveis no intervalo temporal, as regras são armazenadas em uma estrutura própria desenvolvida para manter todas as informações relacionadas à regra. Essa estrutura contém os padrões antecedentes e consequentes, a característica temporal, as métricas de avaliação e as ocorrências nas séries temporais, com seus índices e tempos de início (linhas 79 a 81 do Algoritmo 1).

As regras podem ser retornadas no formato curto ou extenso. Um exemplo de regra temporal no formato curto é $([IMP,I] \Rightarrow [PIB,I], \Delta t = 0)$ com o antecedente sendo um aumento (símbolo I) na variável importação, o consequente sendo um aumento na variável PIB e a característica temporal igual a 0, indicando que o consequente ocorre junto com o antecedente. No formato extenso inclui-se também as ocorrências, por exemplo, (bra,1997;bra,2000;bra,2001), indicando que a regra ocorreu no Brasil (índice “bra”) em 1997, 2000 e 2001.

3.5 Avaliação

As métricas de avaliação mais utilizadas na mineração de regras são suporte e confiança, detalhados na Seção 2.1. No contexto de regras de associação, o suporte pode atingir até o valor 100, mas o eTRUMiner avalia o suporte sobre o conjunto total de regras geradas, que é dividido entre os múltiplos intervalos temporais. Dessa forma, o suporte de uma regra temporal gerada pelo eTRUMiner atinge faixas inferiores, conforme demonstrado na Seção 3.1.

A etapa de avaliação das regras é realizada no final do algoritmo após a composição das regras (Algoritmo 1 linhas 83 - 85). Embora o cálculo das métricas seja feito apenas no final, durante a execução do algoritmo utiliza-se o suporte mínimo para reduzir o volume de regras geradas em cada etapa, evitando o excesso de trabalho computacional para regras com suporte abaixo do mínimo, com base no algoritmo *Apriori*.

O cálculo do suporte realizado no algoritmo utiliza a frequência de ocorrência da intersecção dos padrões, já determinado na etapa de geração de regras, e o total de transações gerado do conjunto de dados, calculado após a etapa de geração de transações (Equação 3.1).

Para calcular a confiança da regra é necessário obter a frequência de transações que contenham todos os padrões antecedentes da regra, frA. A confiança é dada pela divisão dessa frequência sobre o total de transações da regra (Equação 3.3).

As regras retornadas encontram-se acima do suporte mínimo e confiança mínima pré-determinados, e sua ordenação é realizada primeiro pelo suporte e depois pela confiança. O formato retornado pode ser curto, composto pela regra e valores de suporte e confiança, ou extenso, que contém todas as informações do formato curto e as ocorrências nas séries temporais. O algoritmo permite inserir novas métricas de avaliação e também modificar as métricas já contempladas, permitindo diversas formas de avaliação das regras temporais.

3.6 Implementação do eTRUMiner

A implementação do eTRUMiner foi realizada em C++ utilizando o conceito de classes. A escolha de uma linguagem de baixo nível deve-se à característica da mineração de regras temporais, uma tarefa que requer melhor uso da memória por gerar um grande volume de dados durante o processo e que apresenta potencial carga computacional devido à sua característica fatorial de geração de dados. Além disso, com o uso de classes é possível reutilizar e aprimorar o código mais facilmente, simplificando a sua compreensão.

O conjunto de dados a ser analisado pelo eTRUMiner deve ser composto por arquivos de séries temporais univariadas em formato *csv* e podem possuir fontes distintas. Para construção do conjunto de dados, é necessário apenas que o identificador de cada série temporal seja consistente (por exemplo, no conjunto de dados avaliado, as séries do Brasil são referidas como “bra” em todas as variáveis), mantendo exatamente a mesma denominação entre diferentes arquivos para cada variável da mesma série. A ordem de organização entre as séries é irrelevante, facilitando a integração das variáveis de diferentes origens em um único conjunto de séries multivariadas.

Conforme apresentado no Algoritmo 1 linhas 1 - 5, a discretização é realizada por variável, para cada série univariada, permitindo inclusive diferentes discretizações entre variáveis, mas essa análise não foi realizada neste trabalho. O algoritmo permite a fácil inserção de novos métodos de discretização, possibilitando aplicação sobre conjuntos de dados de diferentes áreas de aplicação. A implementação da discretização SAX utiliza uma tabela contendo a faixa de cobertura de cada símbolo, armazenada no arquivo *SAX_dots.csv*. O número máximo de elementos distintos na discretização SAX é 26 que constitui o tamanho do alfabeto.

As principais estruturas de dados utilizadas no eTRUMiner são vetores e dicionários. Por exemplo, o armazenamento dos padrões de cada série é realizado em um dicionário, sendo que cada valor é um vetor de padrões que possuem o mesmo tempo inicial. As transações também são armazenadas em um dicionário, com a chave armazenada sendo a diferença entre os tempos iniciais do conjunto antecedente, e o conjunto de consequente, além de ambos os tempos iniciais. O valor da chave é uma tupla de vetores contendo o conjunto de padrões antecedente e o conjunto

de padrões consequente. Para cada fator temporal Δt, armazena-se um dicionário de padrões que contém um vetor de índices das transações de ocorrência no conjunto antecedente e no conjunto consequente.

3.7 Análise de complexidade

A complexidade de memória depende do conjunto de dados, da discretização utilizada e do número de regras distintas geradas durante o processo de mineração. Para o conjunto de dados, o uso é dado por $\delta.N.n + \delta.n + N$, em que $\delta.N.n$ refere-se as observações armazenadas, $\delta.n$ para manter o tempo das observações univariadas, e N para armazenar os identificadores das séries. O uso de memória para armazenar as séries discretizadas é dado por $\delta.(3.N.L)$, com a constante 3 referindo-se aos componentes do elemento discretizado (símbolo, tempo inicial e tempo final).

Na etapa de geração de transações, para o armazenamento de *ptrn* utiliza-se $\delta.N.L$, e são geradas $N.(w+1).(2L-w)/2$ transações (*seq*). Esses espaços são usados temporariamente para gerar os padrões frequentes que são utilizados durante todo o processo de mineração de regras. O armazenamento dos padrões frequentes e suas ocorrências ocupa $N.(w+1).(2L-w)/2$ para guardar as informações das transações e x padrões frequentes, com x variando de $\delta.(w+1)$ até $\delta.(w+1).(2L-w)/2$. O total de memória utilizado nessa etapa então é dado por $N.(w+1).(2L-w)+x$, em que o índice de cada ocorrência aparece para antecedentes e consequentes.

No pior caso possível são geradas $3^{\delta} - 3.2^{\delta} + 3$ regras para cada transação, sem considerar o corte no suporte. Por isso, o corte de suporte mínimo auxilia na redução da complexidade temporal e espacial, evitando o processamento e o armazenamento de regras sem interesse. Além disso, como o número de transações geradas é reduzido pelo corte da janela temporal, o uso da janela temporal também auxilia na redução de ambas as complexidades, a de tempo de processamento e a de memória.

Para cada regra acima do suporte e da confiança mínimos são armazenados os padrões do antecedente e do consequente, a característica temporal e as ocorrências nas séries temporais. As ocorrências nas séries são armazenadas para cada padrão, mantendo o identificador da série, tempo inicial e tempo final da ocorrência do padrão. O comportamento da memória utilizada é dado por $O(3^{\delta}.N.L.w)$.

Para reduzir a complexidade temporal de geração de regras multivariadas, além do uso da janela temporal, definiu-se uma ordem específica entre variáveis (para formação de regras) e utilizou-se a classe de *map* no armazenamento de padrões e regras. A complexidade da discretização depende do método utilizado, mas dentre os métodos já implementados, a discretização possui complexidade linear com o número de observações das séries, resultando em O($\delta.N.n$).

A etapa de geração de transações ocorre para cada série, de forma que todas as operações são realizadas N vezes. A complexidade dessa etapa no pior caso é dada por $N.[\delta.L.\log(\delta.L) + (\delta.L)^2.log(\delta.L)^2 + (w+1).(\delta.L + 2.\delta.L.\log(\delta.L))]$, em que o primeiro termo é a geração dos padrões, o termo quadrático refere-se à geração de transações e $(w+1).(\delta.L + 2.\delta.L.\log(\delta.L))$ é o armazenamento dos padrões e suas ocorrências.

A geração de regras é realizada para cada característica temporal até atingir a janela temporal, de modo que o processo repete-se $w+1$ vezes. O número de padrões distintos também é consequência do método de discretização utilizado e podem ser gerados até $(w+1).\delta.L.N$ padrões distintos. Considerando esse pior caso e sem nenhum corte de suporte mínimo, a geração de regras possui complexidade $O(w.\delta.3^{\delta}.(3\delta + L.\delta^3 + \delta^3.L^2))$, confirmando-se como a etapa mais custosa. O uso da janela temporal reduz a complexidade de um fator N para w.

A busca das ocorrências de cada regra é realizada para cada padrão que a compõe e associada a todas as ocorrências nas séries. Por tanto, a complexidade possui comportamento $O(3^{\delta}.\delta.\log(N.L))$. A avaliação de suporte e confiança apresenta complexidade $3^{\delta}.\delta^2.L$ enquanto a ordenação é realizada em $O(3^{2.\delta})$.

A complexidade temporal do eTRUMiner no pior caso possível é $O(w.3^{\delta}.\delta^4.L^2)$. Contudo, o algoritmo apresenta algumas heurísticas empregadas para reduzir essa complexidade que não entram na avaliação de pior caso. A definição de ordem dos padrões dentro do antecedente e do consequente, a obrigatoriedade de todos os padrões no conjunto do antecedente ou do consequente possuírem o mesmo tempo inicial e o suporte mínimo reduzem consideravelmente o número de regras geradas e processadas.

3.8 Considerações Finais

O eTRUMiner minera regras temporais multivariadas de séries temporais multivariadas definida a partir de dados de fontes distintas. O algoritmo possui capacidade de lidar com séries heterogêneas e com observações faltantes, sem a necessidade de pré-processamento, e aceita múltiplas discretizações. Os conceitos utilizados para o desenvolvimento dessa solução foram apresentados e discutidos.

O algoritmo divide-se em quatro etapas principais: discretização, geração de transações, geração de regras e avaliação das regras. Na etapa de discretização, as séries temporais são transformadas em séries discretizadas compostas de elementos discretizados e existem quatro métodos já implementados no eTRUMiner: comportamental, decis, quartis e SAX. Durante a geração de transações, os elementos discretizados são transformados em padrões e associados entre si em cada série formando as transações. A partir das transações as regras temporais multivariadas são geradas e na etapa final, avaliadas.

O próximo capítulo apresenta os resultados obtidos do eTRUMiner sobre a aplicação em um conjunto de dados econômicos com observações e variáveis faltantes. Avalia-se a performance do algoritmo sobre os dados, o desempenho no tratamento de valores faltantes e explora-se uma análise semântica das regras obtidas.

Algoritmo 1 – eTRUMiner

Entrada: Conjunto de Dados S, Método de Discretização *disc*, Janela Temporal w, Suporte Mínimo sup_{min}, Confiança Mínima $conf_{min}$

Saída: Regras Temporais Multivariadas (curtas or extensas) acima de sup_{min} e $conf_{min}$ ordenadas

1: **para** cada variável *var* em S **faça**
2: **para** cada série univariada $s_i[var]$ **faça**
3: Discretize pelo método *disc* cada série temporal $s_i[var]$ em $s'_i[var]$
4: **fim para**
5: **fim para**
6: **para** cada série discretizada s'_i **faça**
7: **para** cada variável *var* em s'_i **faça**
8: **para** cada elemento discretizado α_{t_i,t_f} de $s'_i[var]$ **faça**
9: Adicione no vetor de $ptrn[t_i]$: $var-\alpha$.
10: **fim para**
11: **fim para**
12: **para** cada tempo inicial t_i em *ptrn* **faça**
13: t $\leftarrow t_i$
14: **para** t até máximo(t_i) **faça**
15: **se** $(t-t_i) \leq w$ **então**
16: Adicione em $seq[\Delta t; t_i, t]$: ($ptrn[t_i]$, $ptrn[t]$)
17: **fim se**
18: t $\leftarrow (t+1)$
19: **fim para**
20: **fim para**
21: **para** cada elemento e em *seq* **faça**
22: Sendo $seq[e] = (ptrn_{antec}, ptrn_{conseq})$ com $e = \Delta t_e; t_{antec}, t_{conseq}$
23: $x \leftarrow 0$
24: Adicione em $itemset\_info[x]$: $index[s'_e] - \Delta t_e; t_{antec}, t_{conseq}$
25: **para** cada padrão $(var-\alpha)$ em $ptrn_{antec}$ **faça**
26: Adicione no vetor do primeiro item de $patt[\Delta t][var-\alpha]$: x
27: **fim para**
28: **para** cada padrão $(var-\alpha)$ em $ptrn_{conseq}$ **faça**
29: Adicione no vetor do segundo item de $patt[\Delta t][var-\alpha]$: x
30: **fim para**
31: $x \leftarrow (x+1)$
32: **fim para**
33: **fim para**
34: total transações T $\leftarrow$ tamanho(*itemset_info*)
35: **para** cada Δt em *patt* **faça**
36: **para** cada padrão $var-\alpha$ em $patt[\Delta t]$ **faça**
37: **se** tamanho($patt[\Delta t][var-\alpha].primeiro$) $\geq sup_{min}$* T **então**
38: Adicione em *freq_antec*: $var-\alpha$
39: **fim se**
40: **se** tamanho($patt[\Delta t][var-\alpha].segundo$) $\geq sup_{min}$* T **então**
41: Adicione em *freq_conseq*: $var-\alpha$
42: **fim se**
43: **fim para**

para cada padrão $var_a - \alpha_a$ em *freq_antec* **faça**
  para cada padrão $var_c - \alpha_c$ em *freq_conseq* **faça**
    se $var_a \neq var_c$ **então**
      se (ocorrencias $(var_a - \alpha_a) \cap$ ocorrencias $(var_c - \alpha_c)) \geq sup_{min} *$ T **então**
        Adicione em $SRules[var_a - \alpha_a; var_c - \alpha_c]$: [$\cap$ ocorrencias]
        Adicione em $rules[var_a - \alpha_a; var_c - \alpha_c]$: [$\cap$ ocorrencias]
      fim se
    fim se
  fim para
fim para
se numero de variáveis $n_{var} > 2$ **então**
  tamanho da regra $tam_{regra} = 2$
  enquanto $tam_{regra} < n_{var}$ & tamanho($rules$)> 0 **faça**
    para cada regra r_i em *rules* **faça**
      para cada padrão $var_a - \alpha_a$ em *freq_antec* **faça**
        se variáveis em $r_i \neq var_a$ **então**
          se (ocorrencias $r_i \cap$ ocorrencias $(var_a - \alpha_a)) \geq sup_{min} *$ T **então**
            Adicione em $SRules[r_i \cup (var_a - \alpha_a)]$: [$\cap$ ocorrencias]
            Adicione em $new_{rules}[r_i \cup (var_a - \alpha_a)]$: [$\cap$ ocorrencias]
          fim se
        fim se
      fim para
      para cada padrão $var_c - \alpha_c$ em *freq_conseq* **faça**
        se variáveis $r_i \neq var_c$ **então**
          se (ocorrencias $r_i \cap$ ocorrencias $var_c - \alpha_c) \geq sup_{min} *$ T **então**
            Adicione em $SRules[r_i \cup (var_c - \alpha_c)]$: [$\cap$ ocorrencias]
            Adicione em $new_{rules}[r_i \cup (var_c - \alpha_c)]$: [$\cap$ ocorrencias]
          fim se
        fim se
      fim para
    fim para
    $rules \leftarrow new\_rules$
    $tam_{regra} \leftarrow (tam_{regra} + 1)$
  fim enquanto
fim se
para cada regra r_i em *SRules* **faça**
  Adiciona ocorrências dos antecedentes e consequentes
fim para
fim para
para cada regra r_i gerada **faça**
  Avalia suporte *sup* e confiança *conf* da regra
fim para
Ordena regras por *sup* e *conf* excluindo regras $conf < conf_{min}$
Retorna regras no formato curto ou extenso

CAPÍTULO

4

ANÁLISE EXPERIMENTAL

Neste capítulo serão apresentados os experimentos realizados para avaliar a solução proposta, o algoritmo eTRUMiner, quanto à mineração sobre o conjunto de dados do comércio internacional. A avaliação divide-se em quatro grupos de experimentos associados à aplicação, dado que, como mencionado no Capítulo 2, os resultados são indissociáveis das características dos dados. Os dois primeiros experimentos avaliam os parâmetros ideais para o conjunto de dados econômico, o terceiro experimento analisa o desempenho do eTRUMiner sobre conjuntos com dados faltantes, e o último experimento trata-se de uma análise semântica das regras mineradas sobre os dados do comércio internacional.

O conjunto de dados econômico utilizado contém quatro variáveis do comércio internacional: valores transacionados de importação e exportação, Índice de Complexidade Econômica e Produto Interno Bruto. As séries temporais do conjunto de dados são multivariadas, heterogêneas (com duração distinta entre as variáveis) e incompletas (contendo observações e variáveis faltantes).

A janela temporal máxima definida para geração de regras temporais é de $w = 5$. Por se tratar de séries anuais, esse valor implica que o consequente pode ocorrer até 5 anos após o antecedente da regra. Esse valor é usualmente definido com o auxílio do especialista da área de aplicação e indica o tempo máximo que ainda faz sentido semântico para minerar regras. No cenário econômico, o intervalo de 5 anos é um período suficiente quando se analisa os ciclos econômicos[1] (ZARNOWITZ; OZYILDIRIM, 2006).

As discretizações implementadas no eTRUMiner e avaliadas experimentalmente são a comportamental, decis, quartis e SAX (com o número de elementos discretizados sendo 3). Nos experimentos, as discretizações comportamental, decis e quartis são aplicadas sobre os valores absolutos, enquanto a discretização SAX já possui a normalização *z-score* embutida

[1] EABCN <https://eabcn.org/dc/chronology-euro-area-business-cycles>, NBER <https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions>

no método. Para a avaliação das regras temporais, utiliza-se as medidas de suporte e confiança como definidas na Equação 3.1 e na Equação 3.3. Os experimentos realizados e seus respectivos objetivos são:

1. **Análise das Discretizações (Seção 4.2):** Análise das discretizações implementadas (comportamental, decis, quartis e SAX) sobre o conjunto de dados do comércio internacional para avaliação da influência de cada método na geração de regras relevantes.

2. **Análise das Medidas de Corte (Seção 4.3):** Avaliação da distribuição das regras temporais considerando suporte e confiança após aplicação de medidas de corte para escolha quantitativa de valores adequados.

3. **Análise sobre Dados Faltantes (Seção 4.4):** Avaliação comparativa entre conjuntos de dados incompleto e completo para análise do desempenho do eTRUMiner no tratamento de dados faltantes e respectivo impacto nos resultados.

4. **Análise Semântica (Seção 4.5):** Análise semântica das regras mineradas sobre o conjunto de dados do comércio internacional para avaliar a confiabilidade e a coerência das regras.

A seguir, apresenta-se o conjunto de dados utilizado e avaliado, composto de séries econômicas do comércio internacional. Os experimentos realizados são detalhados nas seções seguintes, com a análise das discretizações na Seção 4.2, análise das medidas de corte na Seção 4.3, análise sobre dados faltantes na Seção 4.4 e análise semântica na Seção 4.5. A Seção 4.6 apresenta as considerações finais.

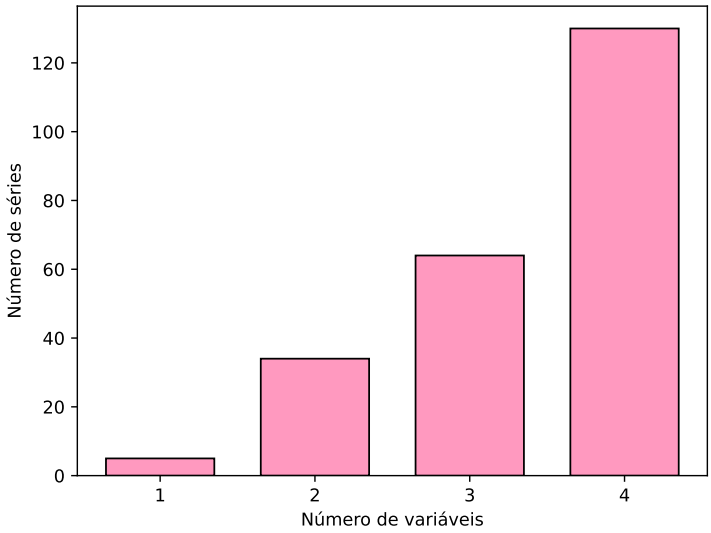
4.1 Conjunto de Dados

O conjunto de dados utilizado para avaliação do eTRUMiner constitui-se de séries temporais multivariadas, com observações faltantes, variáveis ausentes em algumas séries e duração distinta entre variáveis. Ele foi construído a partir da integração de dados do comércio internacional de países e territórios, que por simplicidade serão denominados apenas "países", abrangendo valores transacionados de importação (IMP) e exportação (EXP), Índice de Complexidade Econômica (ECI) e o Produto Interno Bruto (PIB). Essas séries possuem periodicidade anual e são fornecidas por fontes variadas.

A extração do conjunto de dados foi realizada em março de 2022, totalizando 232 séries contendo até 4 variáveis no conjunto de dados denominado como "original". Cada série possui um identificador de 3 letras que segue o ISO 3166[2] e refere-se a um país distinto. Considerando o conjunto de 232 séries, com 4 variáveis e 25 anos de duração, o percentual geral de observações faltantes do conjunto de dados original é de $9,39\%$, contendo apenas 5 séries univariadas.

[2] ISO 3166 <https://www.iso.org/iso-3166-country-codes.html>

Figura 11 – Distribuição das séries pelo seu número de variáveis (até 4 variáveis abrangendo importação, exportação, ECI e PIB) com cada série referindo-se a um país.



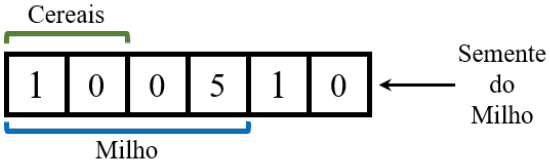
Fonte: Elaborada pela autora.

A distribuição do número de variáveis entre as séries é apresentada na Figura 11, com 130 séries possuindo todas as variáveis. Essa quantidade representa apenas 56,03% do conjunto de países avaliados. Para melhor compreensão dos dados, Subseção 4.1.1 a Subseção 4.1.3 contêm mais detalhes sobre cada variável econômica utilizada.

4.1.1 Importação e Exportação

O conjunto de dados de importação e exportação é derivado do CEPII[3], o principal centro francês de pesquisa e expertise em economia internacional. A base contém os valores anuais do comércio de mercadorias de 228 países de 1996 a 2020, abrangendo mais de 5.000 produtos detalhados através da nomenclatura de 6-dígitos *Sistema Harmonizado 96*. Cada registro contém o país de origem, de destino, código do produto e valor da transação em milhares de dólares.

Figura 12 – Descrição da composição do código numérico de um produto conforme a nomenclatura *Sistema Harmonizado 96*.



Fonte: Elaborada pela autora.

O produto transacionado é representado por um código numérico (denominado *hs96*) contendo 6 dígitos. Os dois primeiros dígitos indicam o grupo, o conjunto dos quatro primeiros

[3] BACI <http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=37>

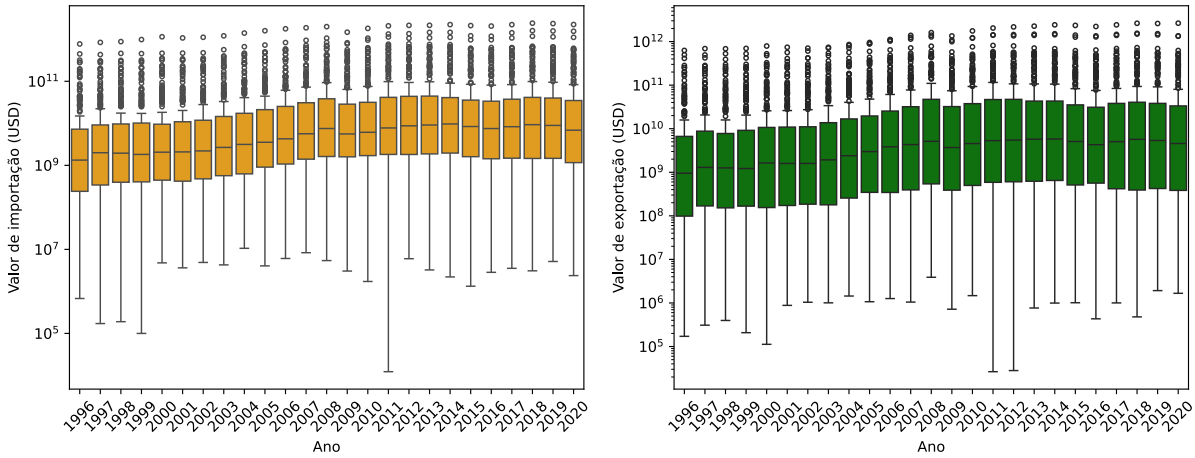
indicam o produto, enquanto os dois últimos permitem uma descrição mais detalhada a respeito do item. Por exemplo, o grupo 10 abrange cereais, 1005 refere-se a milho e o código 100510 refere-se à semente do milho, conforme ilustrado na Figura 12.

Embora o conjunto permita uma análise aprofundada referente aos valores transacionados para cada produto específico, o foco deste trabalho está apenas nos montantes totais transacionados entre os países. Para a obtenção dos valores anuais de exportação de cada país, agrupou-se cada transação por origem no ano, somando-se os valores transacionados. Para a importação, o agrupamento é sobre o destino no ano.

Figura 13 – Variação anual de 1996 a 2020 dos valores de importação e exportação do conjunto de dados original.

(a) Importação em dólares.

(b) Exportação em dólares.

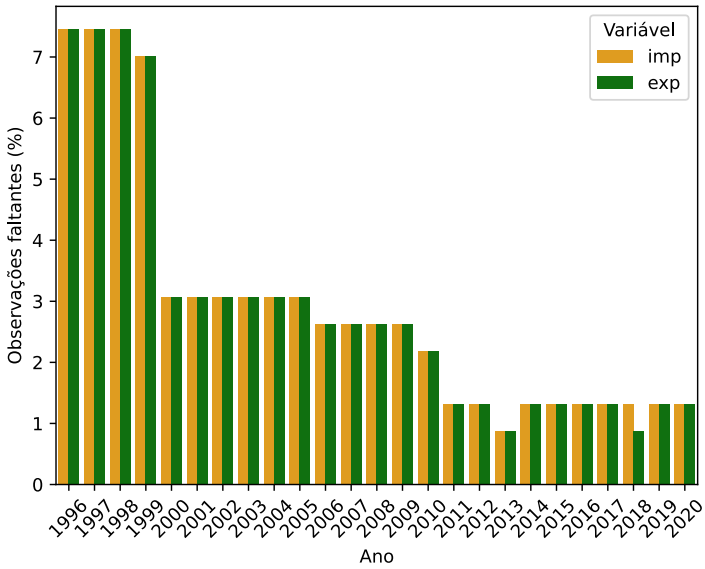


Fonte: Elaborada pela autora.

A distribuição anual dos valores de importação de produtos entre países é apresentada na Figura 13a, enquanto a Figura 13b apresenta a distribuição anual dos valores de exportação. Ambos os gráficos possuem tendência de crescimento, mas verifica-se uma dispersão maior nos valores anuais de exportação. É possível observar os reflexos de duas crises econômicas com a distribuição de valores atingindo os mínimos do período tanto no gráfico de importação quanto no de exportação, em 1998-1999 (Crise da Bolha de Internet) e 2010-2012 (Grande Recessão Econômica de 2009). A partir do ano de 2019 há uma tendência de queda nos valores de importação e exportação, indicando a contração das economias conforme constatado em UNCTAD (2020)

As séries de importação do conjunto de dados original possuem 167 observações faltantes, enquanto na exportação o número total de observações faltantes é 166, representando menos de 3% do total de observações em cada uma. A Figura 14 apresenta a distribuição anual percentual de observações faltantes nas variáveis de importação (*imp*) e exportação (*exp*).

Figura 14 – Distribuição anual de 1996 a 2020 das observações faltantes do conjunto de dados original nas variáveis importação e exportação.



Fonte: Elaborada pela autora.

Os primeiros anos das séries (dados mais antigos) possuem os maiores percentuais de observações faltantes. Isso ocorre devido ao processo coleta, que é declaratório e a obrigatoriedade de divulgação foi gradualmente abrangendo os países ao longo dos anos. A partir de 2010 a quantidade anual de dados faltantes estabilizou-se abaixo de 2%.

4.1.2 Índice de Complexidade Econômica

O Índice de Complexidade Econômica é fornecido pelo Observatório de Crescimento de Harvard através do Atlas de Complexidade Econômica [4] e indica a diversidade e a complexidade da capacidade produtiva de um país com base nos produtos exportados. Os valores variam de -3 a 3 e quanto maior o valor, maior a capacidade do país de produzir um conjunto altamente diversificado de produtos com alta complexidade.

As séries possuem periodicidade anual e abrangem 132 países de 1996 a 2019. A distribuição anual dos valores de ECI podem ser visualizados na Figura 15a. Enquanto a média dos valores anuais não apresenta grande alteração com o passar dos anos, permanecendo por volta do 0, é possível observar um aumento na dispersão com o tempo. Isso demonstra um crescimento na disparidade da capacidade produtiva entre os países, problema discutido pelas Nações Unidas e o WTO (*World Trade Organization*) (UNCTAD, 2023).

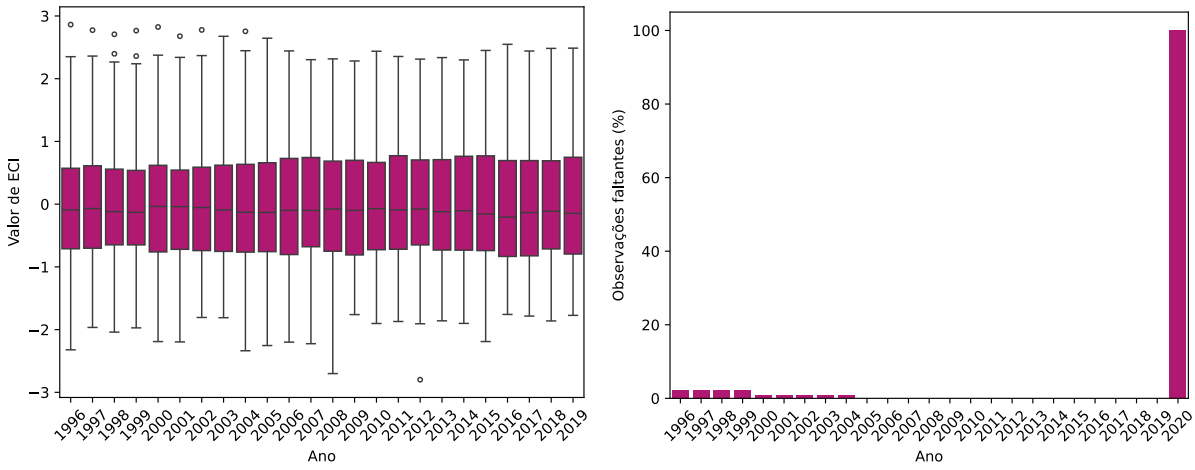
A Figura 15b apresenta a distribuição anual percentual de observações faltantes na variável de ECI, que representa menos de 1% do total de observações. Entre 1996 e 1999, o percentual de observações faltantes manteve-se no patamar de 3%, referente a 3 países do

[4] ECI <https://atlas.cid.harvard.edu/rankings>

Figura 15 – Distribuição anual de valores (1996 a 2019) e percentual de observações faltantes (1996 a 2020) da variável ECI no conjunto de dados original.

(a) Variação anual do ECI entre países.

(b) Distribuição anual de valores faltantes no ECI.



Fonte: Elaborada pela autora.

conjunto sem ECI. A partir de 2000 e nos 4 anos seguintes, apenas a Sérvia mantém-se sem observação, representando menos de 1% do conjunto de séries. Ao todo, existem 17 observações faltantes até 2019 e o ano de 2020 é representado com 100% de observações faltantes para todas as séries pois os dados ainda não estavam disponíveis no momento da extração. No histograma (Figura 15b), a coluna 2020 é apresentada pois trata-se do último ano presente nas demais variáveis.

4.1.3 Produto Interno Bruto

O Fundo Monetário Internacional (FMI) [5] fornece anualmente o Produto Interno Bruto dos países em bilhões de dólares. As séries temporais extraídas para análise cobrem o PIB anual de 1996 a 2020 de 196 países, com um total de 146 observações faltantes. A tendência de crescimento verificada na Figura 16a é o comportamento esperado do PIB, como apresentado em WorldBank e OECD (2023), e a grande dispersão dos valores deve-se à enorme disparidade econômica existente entre os vários países avaliados.

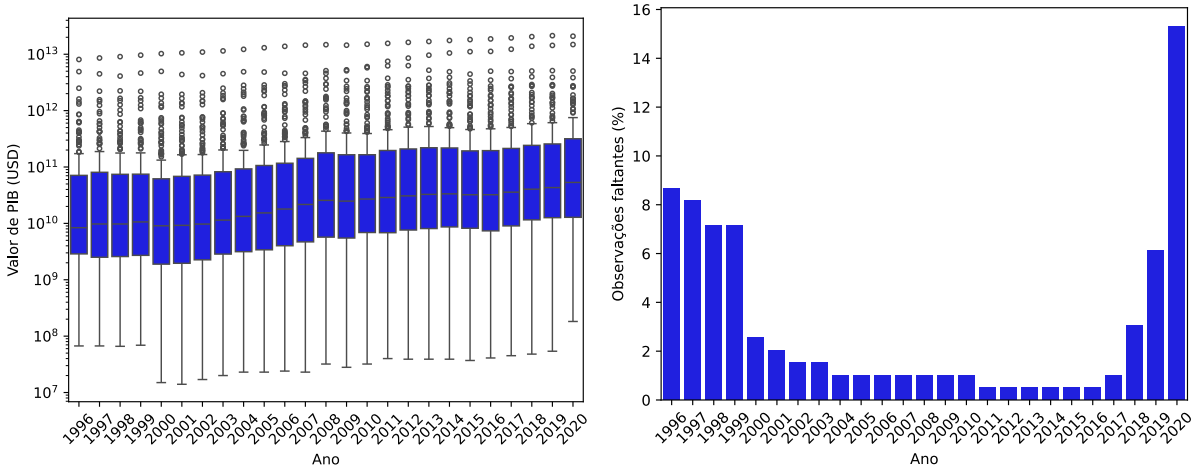
A Figura 16b apresenta a distribuição anual percentual de observações faltantes no PIB. No total de observações dessa variável, o percentual faltante representa menos de 3% e, dado que o PIB é fornecido pelo seu respectivo governo, é esperado que exista uma parcela de observações faltantes. O período de maior disponibilidade de informações está entre 2011 e 2016, com crescente quantidade de observações faltantes nos anos seguintes. Esse comportamento pode ser devido a atrasos na disponibilização das informações pelos governos para os órgãos de controle.

[5] GDP <https://www.imf.org/en/Publications/WEO/weo-database/2022/April>

Figura 16 – Distribuição anual (1996 a 2020) de valores e percentual de observações faltantes da variável PIB no conjunto de dados original.

(a) Variação anual do PIB entre países em dólares.

(b) Distribuição anual de valores faltantes no PIB.



Fonte: Elaborada pela autora.

4.1.4 Conjunto de Dados Homogêneo

A partir do conjunto de dados original, selecionou-se as séries completas (sem variáveis faltantes e sem observações faltantes) e gerou-se um conjunto de dados reduzido mantendo as variáveis importação, exportação, ECI e PIB. Esse conjunto de dados denominado "homogêneo" é utilizado para análise e comparação de resultados em relação ao conjunto original, visando avaliar a capacidade do eTRUMiner de lidar com dados faltantes, e para análise semântica.

O conjunto de dados homogêneo compõe-se de 116 países que representam cerca de metade das séries presentes no conjunto original. A Figura 17 apresenta a distribuição das variáveis no conjunto de dados original e no homogêneo, mostrando que em nível comportamental trata-se de uma amostra adequada com menor variabilidade. Como a variável ECI não apresenta observações no ano de 2020, para todas as variáveis do conjunto de dados homogêneo as séries são limitadas até 2019.

O Apêndice A apresenta os países e territórios contidos no conjunto de dados original, bem como as séries presentes no conjunto de dados homogêneo, em negrito. A primeira coluna apresenta o nome em português do país ou território com a sua respectiva sigla ISO 3166 na segunda coluna, que é o identificador das séries.

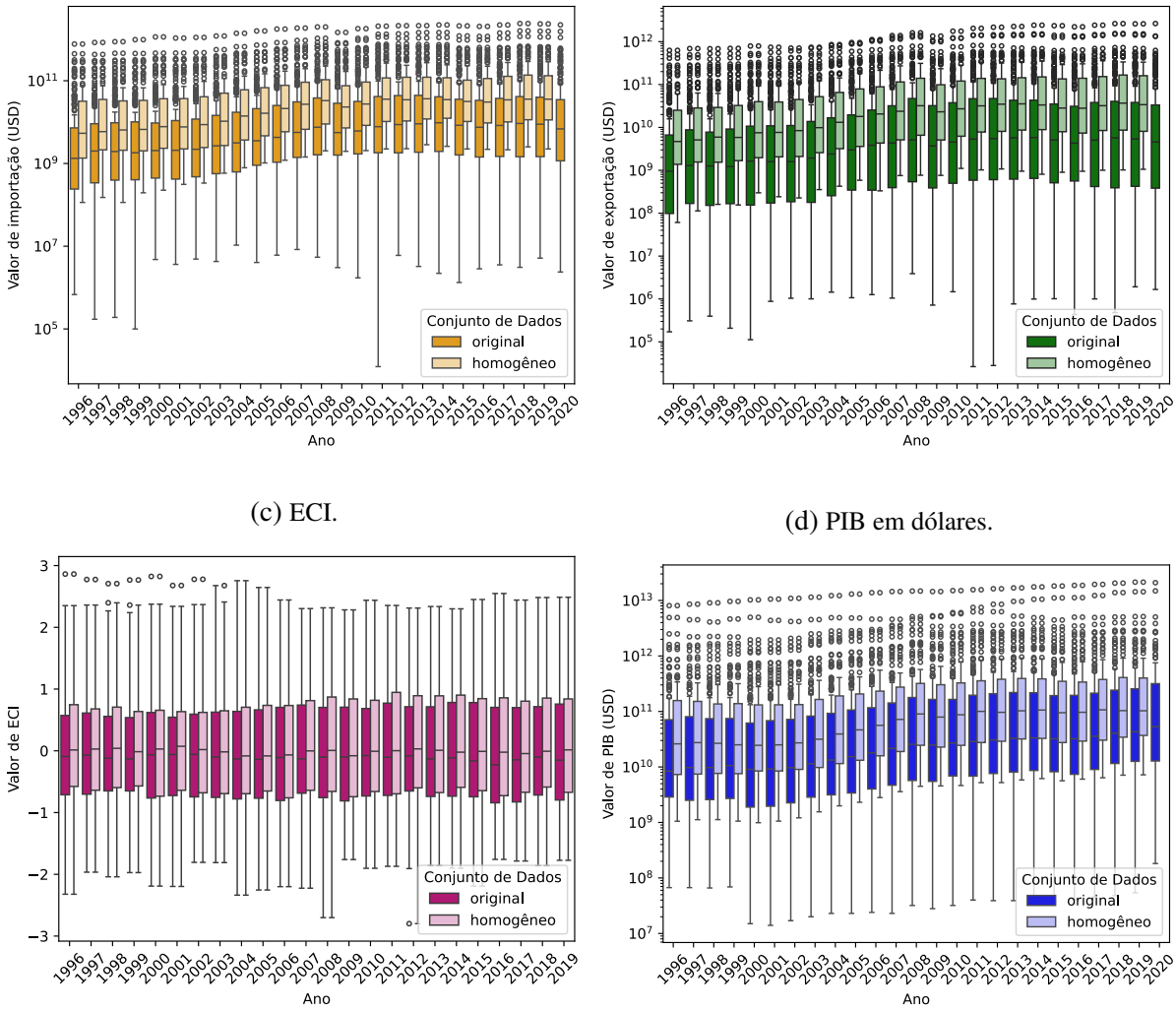
4.2 Análise das Discretizações

Esta seção contém os experimentos realizados para avaliar os métodos de discretização implementados, selecionados com base nas características do conjunto de dados utilizado. As discretizações comportamental, decis e quartis foram aplicadas sobre os valores absolutos das

Figura 17 – Comparativo da distribuição anual de valores entre conjunto de dados por variável.

(a) Importação em dólares.

(b) Exportação em dólares.



(c) ECI.

(d) PIB em dólares.

Fonte: Elaborada pela autora.

séries do conjunto de dados original para evitar a geração de alterações devido à manipulação de dados. Contudo, na discretização SAX aplica-se a normalização *z-score* sobre o conjunto de dados original como requisito do método.

As discretizações comportamental e SAX, que geram três elementos discretizados distintos, retornam quantidades semelhantes de regras, na ordem de dez mil regras distintas. Contudo, as discretizações decis e quartis retornam centenas de milhares de regras distintas, indicando que o número de elementos discretizados distintos que podem ser gerados pelo método de discretização possui forte influência sobre o número de regras temporais distintas geradas.

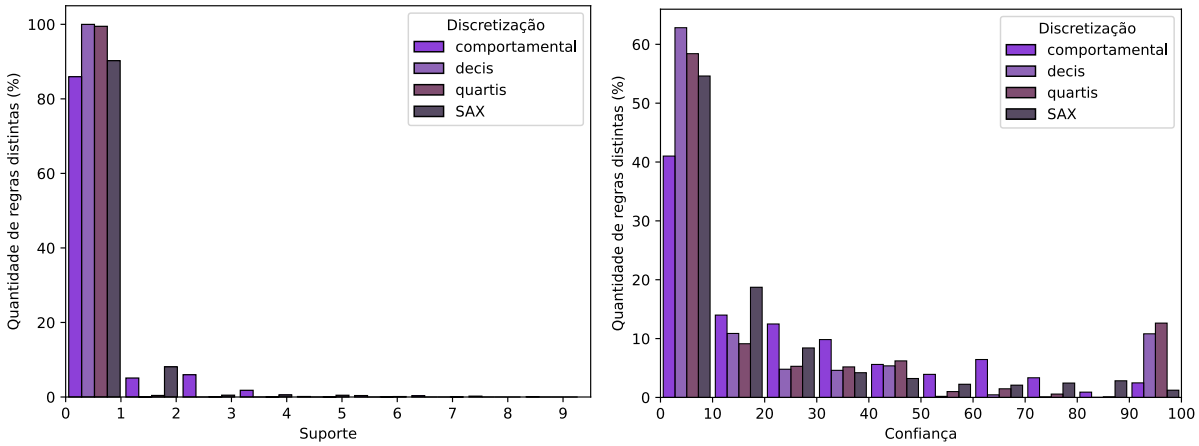
A discretização comportamental, além de gerar o menor volume de regras distintas, apresenta o maior valor de suporte máximo entre todos os métodos de discretização avaliados, indicando uma menor dispersão existente entre as regras geradas. Entre as discretizações decis e

quartis, a discretização quartis apresenta maior suporte máximo, e o número de regras distintas geradas sobre o conjunto de dados original é menor que a metade do volume gerado pela discretização decis.

A avaliação detalhada dos métodos de discretização é realizada através da distribuição de regras no suporte e na confiança, as métricas de avaliação utilizadas. A Figura 18 apresenta a distribuição percentual das regras entre os valores de suporte (Figura 18a) e confiança (Figura 18b) para cada método de discretização sobre o conjunto de dados original. Contudo, os valores de confiança percentuais não são diretamente comparáveis entre os métodos de discretização. Os valores são apresentados em percentuais pois a quantidade total de regras é proporcional ao número de elementos discretizados distintos que cada discretização permite gerar, conforme detalhado na Subseção 2.2.2.

Figura 18 – Distribuição percentual das regras temporais nas métricas de avaliação para todas as discretizações sobre o conjunto de dados original.

(a) Suporte. (b) Confiança.



Fonte: Elaborada pela autora.

Em todas as discretizações avaliadas, acima de 80% das regras possuem suporte muito baixo, entre 0 e 1. Na discretização comportamental, o suporte máximo atingido é $9,32$, sendo $0,05\%$ do total das regras com suporte 9 entre 10. Na discretização SAX, a quantidade de regras distintas com suporte entre 0 e 1, é de aproximadamente 90%, e o suporte máximo atinge $6,29$. O perfil da distribuição no suporte entre as regras se assemelha à discretização comportamental.

A discretização decis é o método que produz mais regras distintas entre as implementadas, com mais de 95% das regras apresentando suporte até $0,1$. Na discretização quartis, as regras com suporte de até $0,5$ abrangem mais de 98% do total de regras geradas. O método quartis é uma discretização mais informativa que a comportamental, mas que gera um grande volume de regras distintas, embora menor que a decis, dificultando a análise de seus resultados.

A Figura 18b contém a distribuição na confiança para todos os métodos de discretização avaliados. Nas discretizações comportamental e SAX, o pico de regras verificado na faixa de confiança entre 0 e 10, é cerca de 40% das regras na discretização comportamental e acima de 50% na discretização SAX. Para valores de confiança mais altos, o percentual cai exponencialmente, com um novo pico de regras na discretização SAX para confiança entre 70 e 90.

As regras geradas pelas discretizações decis e quartis apresentam um pico de aproximadamente 60% de frequência na menor faixa de confiança (entre 0 e 10) e cerca de 10% da frequência entre a faixa de confiança de 90 a 100. Regras com alta confiança podem ser provenientes de transações com baixa frequência e que consequentemente apresentam um suporte baixo e confiança elevada, ou regras com alto suporte, sendo relevantes. A hipótese é de que tratam-se de regras com baixo suporte e que serão removidas com aplicação de um corte no suporte (sup_{min}), desaparecendo com esse pico.

Dentre as discretizações avaliadas para o conjunto de dados original, o método comportamental apresentou melhores resultados, com menor percentual de regras distintas com baixo suporte (entre 0 e 1) e com baixa confiança (entre 0 e 10). A discretização SAX apresentou resultados similares à discretização comportamental na distribuição no suporte e no volume de regras, mas com suporte máximo mais baixo. Entre as discretizações decis e quartis, que há maior detalhamento nos padrões das regras, o método quartis apresentou resultados melhores, com menor número de regras distintas geradas e menor percentual de regras com baixa confiança.

A alta presença de valores baixos nas métricas de avaliação, suporte entre 0 e 1 e confiança entre 0 e 10, indicam presença de regras não significativas, ressaltando a necessidade da aplicação de medidas de corte para selecionar as regras temporais mais relevantes. A delimitação de suporte mínimo (sup_{min}) e confiança mínima ($conf_{min}$) pode auxiliar a análise do especialista, reduzindo o volume de regras retornadas. Na seção a seguir, apresenta-se uma abordagem quantitativa para escolha desses parâmetros para as discretizações comportamental, quartis e SAX.

4.3 Análise das Medidas de Corte

Em tarefas de mineração de regras, a análise dos resultados e das respectivas métricas de avaliação é realizada com o auxílio de cortes, que determinam os valores mínimos dessas métricas e retornam as regras com valor igual ou acima do mínimo pré-determinado. O valor do corte é usualmente definido pelo especialista da área de aplicação, que possui conhecimento profundo do assunto e é capaz de avaliar semanticamente as regras retornadas. Contudo, como frequentemente ocorre na área computacional, o volume de resultados retornados pode ser muito alto e gerar uma alta carga de trabalho para o avaliador, necessitando de uma abordagem quantitativa para uma avaliação mais adequada.

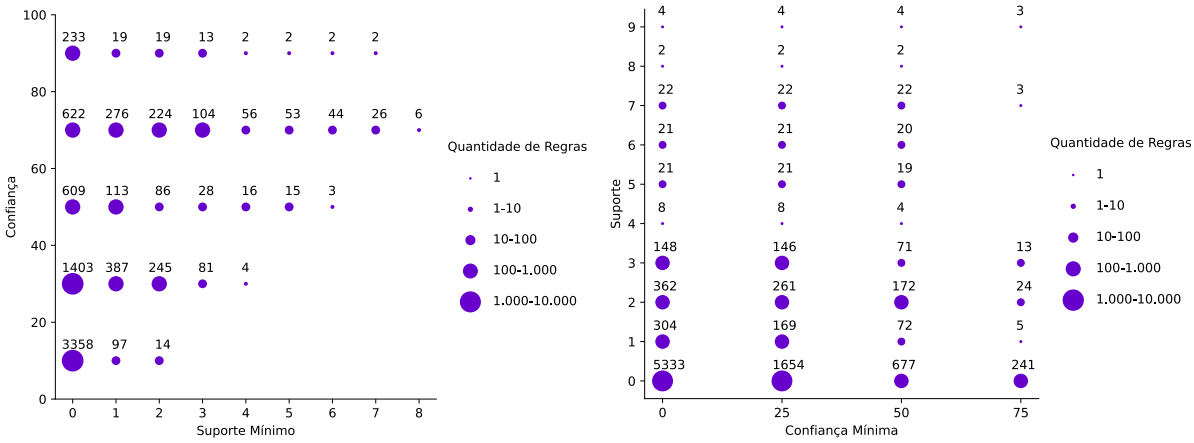
A seguir, apresenta-se uma análise numérica sobre o conjunto de dados original discretizado nos métodos comportamental, quartis e SAX para determinar o corte nas métricas de

avaliação. Essa abordagem quantitativa é independente da semântica dos resultados e permite gerar as regras potencialmente relevantes sem a necessidade do especialista nessa etapa. Realizar essa análise com a posterior com avaliação semântica auxiliada por especialistas pode gerar resultados finais mais eficientemente com elevada acurácia. Para melhor compreensão da influência do corte, a análise está dividida por método de discretização.

Figura 19 – Distribuição das regras após aplicação de corte na **discretização comportamental** para o conjunto de dados original.

(a) Corte no suporte.

(b) Corte na confiança.



Fonte: Elaborada pela autora.

A Figura 19 apresenta, para a discretização comportamental, a distribuição das regras remanescentes após aplicação de corte nas métricas de avaliação. A primeira coluna dos dois gráficos refere-se à distribuição sem corte, $sup_{min} = 0$ e $conf_{min} = 0$, apresentado para fins de referência. A distribuição das regras nas faixas de confiança para aplicação de corte no suporte é apresentada na Figura 19a. Cada linha contém a quantidade de regras na faixa de confiança dos dois marcadores verticais que se encontra, sendo cada ponto a quantidade no corte do suporte.

Por exemplo, o ponto mais à esquerda na primeira linha na Figura 19a é a quantidade de regras com $sup_{min} = 0$ e confiança entre 0 e 20. A confiança até 20 é a faixa com maior número de regras dentro de $sup_{min} = 0$. Isso indica que sem o corte no suporte, o maior número de regras retornadas na discretização comportamental são regras com confiança extremamente baixa, e portanto, irrelevantes.

Para o primeiro corte no suporte, $sup_{min} = 1$, a faixa de confiança com maior número de regras aumenta para 20 a 40. A faixa principal de confiança altera novamente em $sup_{min} = 3$ para 60 a 80 e permanece nessa faixa até o maior corte no suporte avaliado, indicando $sup_{min} \geq 3$ como um corte adequado. Com o corte no suporte $sup_{min} = 8$, do total de regras geradas pelo conjunto de dados original, apenas 6 são retornadas e todas possuem confiança entre 60 a 80. As regras com confiança baixa apresentam tendência de desaparecer com o alto corte no suporte,

indicando a aplicação de corte adequado no suporte como um bom método para a seleção de regras.

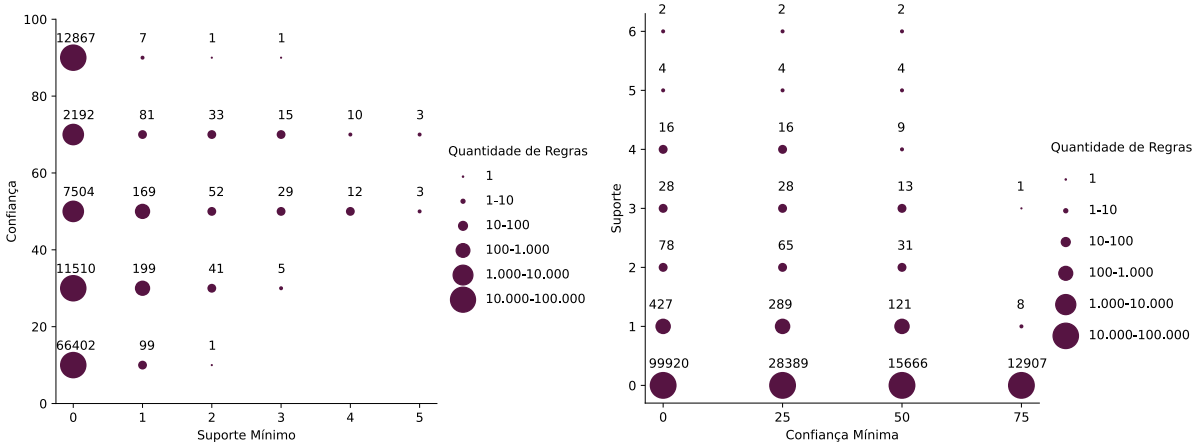
A distribuição das regras nos valores de suporte após corte na confiança é apresentada na Figura 19b, com confiança mínima ($conf_{min}$) variando de 0 a 75. A principal concentração das regras está no suporte entre 0 e 3 para todas as confianças mínimas avaliadas. Para cortes mais altos, regras com suporte até 2 apresentam queda crescente e considerável na quantidade. Para regras com suporte acima de 2, a queda no volume começa em $conf_{min} = 50$ e torna-se significativa apenas para $conf_{min} = 75$, indicando grande resiliência dessas regras para aplicação de corte na confiança. Dado que a redução significativa de volume do regras com suporte alto ocorre após o $conf_{min} = 50$, um corte na confiança nessa faixa é indicada para a discretização comportamental sobre o conjunto original.

A Figura 20 refere-se aos resultados de cortes nas métricas de avaliação aplicados na discretização quartis. Na Figura 20a tem-se a distribuição das regras nos valores de confiança para cortes no suporte (Figura 20a). A distribuição das regras nas faixas de suporte para cortes na confiança é apresentada na Figura 20b.

Figura 20 – Distribuição das regras após aplicação de corte na **discretização quartis** para o conjunto de dados original.

(a) Corte no suporte.

(b) Corte na confiança.



Fonte: Elaborada pela autora.

A discretização quartis gera um alto volume de regras, concentrado na faixa de confiança até 20 para resultados sem o corte. Há um grande pico de regras com confiança muito alta, de 80 a 100, mas esse perfil desaparece logo no primeiro corte do suporte, $sup_{min} = 1$. Esse comportamento mostra que tratam-se de regras com suporte reduzido e portanto de baixa relevância, conforme discussão da Seção 4.2. A alta produção de regras com confiança alta e suporte baixo decorre do método de discretização, que possui uma grande variedade de elementos discretizados distintos e retorna séries discretizadas com baixa coincidência, gerando muitas

regras com baixo suporte. Com o aumento do corte no suporte, verifica-se a mudança na faixa principal de confiança na Figura 20a. Esse aumento estabiliza-se em $sup_{min} = 2$ com a confiança entre 40 e 60.

A quantidade desbalanceada de regras com suporte menor que 1 é verificada para todos os cortes de confiança avaliados, conforme gráfico da Figura 20b. Apenas $12,85\%$ das regras geradas possuem confiança acima de 75, sendo que apenas uma regra possui suporte acima de 1. Além disso, o maior suporte desse corte ($sup = 3$), é aproximadamente metade do suporte máximo atingido pelas regras geradas para o conjunto de dados original com a discretização quartis.

O corte apenas na confiança não é suficiente para selecionar regras relevantes devido ao considerável volume de regras com suporte entre 0 e 1 retornadas. A aplicação de um corte $sup_{min} \geq 2$ em conjunto, pode auxiliar na seleção de algumas regras relevantes, já que a principal faixa de confiança das regras nesse corte de suporte é entre 40 e 60. Contudo, os resultados da análise de corte indicam que a discretização quartis ainda apresenta uma variabilidade muito grande na quantidade de símbolos distintos gerados pelo método, que pode ser deduzido pelo alto volume de regras com suporte e confiança baixos, sendo o seu uso aconselhado para uma mineração de regras mais específica, como por exemplo, para encontrar regras mais detalhadas de regras com padrões mais genéricos já selecionadas.

A Figura 21 apresenta os resultados do corte no suporte e na confiança para a discretização SAX. Na avaliação de corte no suporte, Figura 21a, a faixa de confiança mais comum entre as regras é de 60 a 80, já predominante a partir de $sup_{min} = 2$ e cortes maiores. O corte na confiança apresentado na Figura 21a, mantém grande parte das regras fortes, sendo que na faixa mais alta de suporte a manutenção é de 100% das regras para todos os cortes.

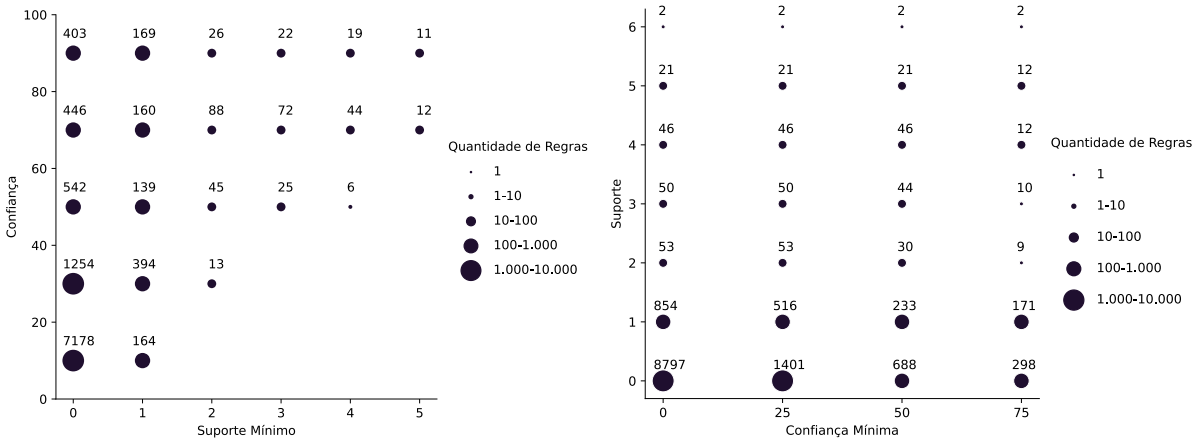
Para o conjunto de dados original discretizado no método SAX, os cortes $sup_{min} \geq 2$ e $conf_{min} \geq 50$ selecionam as regras mais confiáveis. A aplicação do corte $sup_{min} = 2$ e $conf_{min} = 50$ resultam na redução a 1% do total das regras geradas por essa configuração sobre o conjunto original, indicando uma boa seleção para auxiliar o especialista.

A discretização SAX é comumente utilizada e se mostra promissora também na geração de regras temporais a partir de dados econômicos. Entre as discretizações comportamental e SAX, a comportamental fornece regras com suporte mais elevado e a sua interpretação é mais simples por não depender de conhecimento relacionado à distribuição das observações das séries específicas de ocorrência das regras. Em todas as discretizações, as regras com confiança baixa tendem a desaparecer com o corte no suporte, assim como o alto volume de regras com confiança muito alta, indicando a aplicação de corte no suporte como um bom método para a seleção de regras relevantes.

Figura 21 – Distribuição das regras após aplicação de corte na **discretização SAX** para o conjunto de dados original.

(a) Corte no suporte.

(b) Corte na confiança.



Fonte: Elaborada pela autora.

4.4 Análise sobre Dados Faltantes

A presente seção contém os experimentos realizados para avaliar a capacidade do algoritmo, eTRUMiner, de minerar um conjunto de dados com presença de observações e variáveis faltantes. Na Subseção 4.4.1, compara-se os resultados da mineração entre os conjuntos de dados original e homogêneo, avaliando número de regras distintas, suporte máximo obtido e distribuição percentual das regras nas métricas de avaliação. A Subseção 4.4.2 apresenta a análise da influência da presença de observações faltantes no volume de regras temporais geradas e nas métricas de avaliação obtidas.

4.4.1 Comparação dos Conjuntos Original e Homogêneo

A avaliação do desempenho do eTRUMiner entre os conjuntos de dados original e homogêneo é realizada para as discretizações comportamental, quartis e SAX, sem a delimitação de suporte mínimo e confiança mínima. Analisa-se o volume de regras distintas gerado, o suporte máximo atingido e os valores das métricas de avaliação.

Tabela 1 – Número de regras distintas geradas em cada configuração de conjunto de dados e discretização avaliados.

	Original	**Homogêneo**
Comportamental	6.225	6.039
Quartis	100.475	87.043
SAX	9.823	8.978

Fonte: Dados da pesquisa.

A Tabela 1 mostra o total de regras distintas obtidas para as discretizações avaliadas sobre ambos os conjuntos de dados. A variação máxima do número de regras para uma mesma discretização entre os conjuntos é de 15%, com os menores volumes no conjunto homogêneo independentemente da discretização. Dado que trata-se de um subconjunto do conjunto original, estima-se que a menor dispersão de regras é uma consequência da presença de todas as observações e o balanceamento existente entre as variáveis.

As discretizações comportamental e SAX geram o mesmo número de elementos discretizados distintos e apresentam quantidades de regras na mesma ordem de grandeza. Contudo, o volume de regras geradas pela discretização SAX é aproximadamente 50% maior que a quantidade gerada pela discretização comportamental, para ambos os conjuntos de dados. A discretização quartis gera um volume de regras muito maior que as demais, tanto no conjunto original quanto no homogêneo. Embora o método forneça regras mais detalhadas, informando o quartil da variação relativa entre observações consecutivas (Subsubseção 2.2.2.3), o alto volume gerado implica maior dificuldade na análise dos resultados pelo especialista.

Tabela 2 – Suporte máximo obtido em cada configuração de conjunto de dados e discretização.

	Original	**Homogêneo**
Comportamental	9,32	11,84
Quartis	6,12	8,36
SAX	6,29	7,46

Fonte: Dados da pesquisa.

Na Tabela 2, as discretizações quartis e SAX apresentam valores de suporte máximo similares, embora a discretização quartis gere um volume dez vezes maior que a quantidade de regras geradas no método SAX. Com o auxílio da análise de medidas de corte, verifica-se que após a aplicação do corte, o volume de regras fortes retornadas pelos métodos apresenta similaridade. Já a discretização comportamental, além de gerar o menor volume de regras distintas, apresenta os maiores valores de suporte máximo para ambos os conjuntos de dados avaliados, indicando que o método possui boa capacidade de sumarizar os comportamentos existentes nas séries e resulta em grupos de regras com alto volume de ocorrência.

O número de séries avaliado pelo conjunto de dados originais é quase o dobro do conjunto homogêneo, impactando tanto na dispersão das regras geradas quanto no suporte máximo atingido. Contudo, o baixo valor de suporte máximo atingido em todas as configurações, é intrínseco às regras temporais mineradas devido ao alto número de transações gerado (Equação 3.2), conforme detalhado na Seção 3.1.

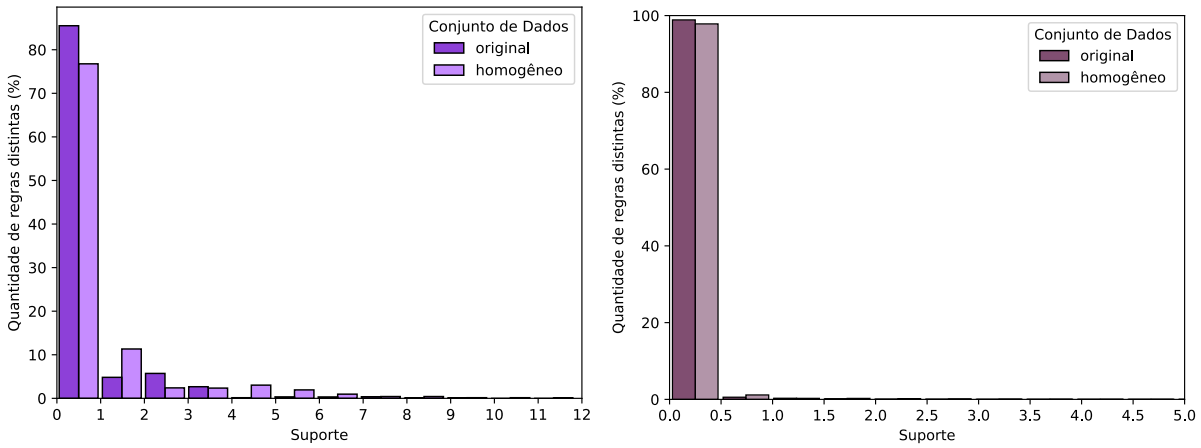
Para melhor visualização dos resultados de cada discretização, a Figura 22 apresenta o gráfico de cada método, comparando os resultados de ambos os conjuntos de dados. Dado que o conjunto homogêneo trata-se de um subconjunto do original, espera-se que o perfil da distribuição do suporte seja semelhante entre conjuntos de dados para as discretizações comportamental e

quartis. A discretização SAX deve ser analisada com maior cautela, já que a remoção do ano de 2020 nas séries do conjunto homogêneo impactam em todos os padrões obtidos para as séries discretizadas, dado que o método considera toda a distribuição para discretizar as observações.

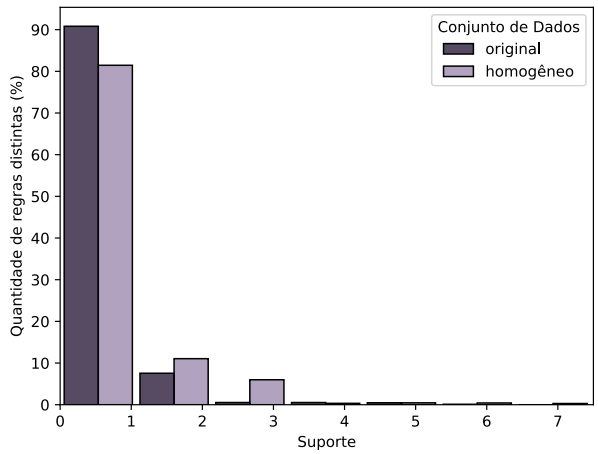
Figura 22 – Distribuição do suporte das regras dividida entre os métodos de discretização comportamental, quartis e SAX para os conjuntos de dados original e homogêneo.

(a) Comportamental.

(b) Quartis.



(c) SAX.



Fonte: Elaborada pela autora.

Na Figura 22a visualiza-se as distribuições para a discretização comportamental. No conjunto de dados original, $85,67\%$ das regras possuem valor de suporte entre 0 e 1, e no conjunto homogêneo, essa faixa do suporte abrange $76,85\%$ das regras. A quantidade de regras no conjunto homogêneo que possui suporte acima do valor máximo obtido no conjunto de dados original, $9,32$, é apenas dezenove, indicando a possibilidade das regras geradas entre os conjuntos possuírem baixa variação no suporte.

A distribuição no suporte das regras geradas pela discretização quartis é apresentada na Figura 22b. Em ambos os conjunto de dados, o percentual de regras geradas que possui suporte

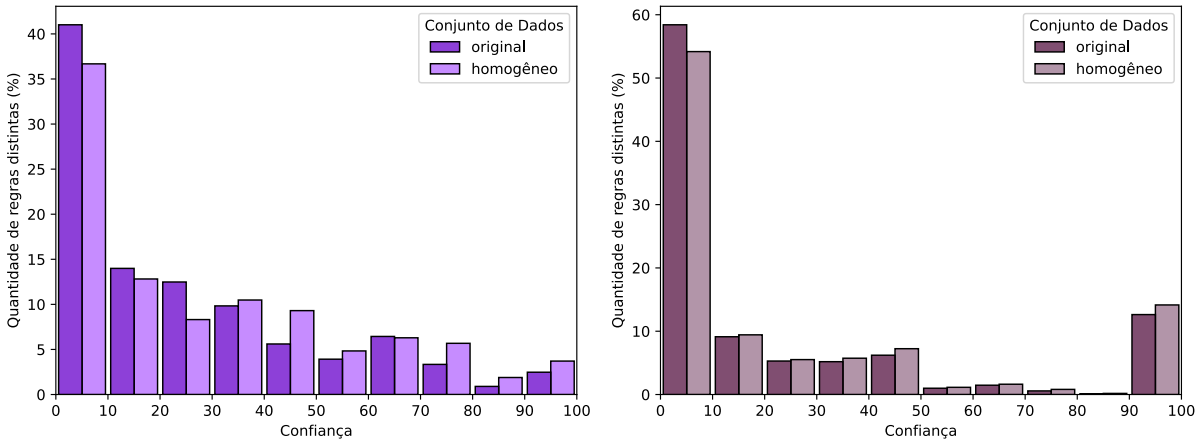
acima de 1 é muito baixo. Além disso, a diferença entre os conjuntos na distribuição percentual é mínima. Trata-se de uma discretização mais informativa que a comportamental, mas que ainda gera um grande volume de regras distintas, dificultando a análise de seus resultados.

Os resultados da discretização SAX entre os dois conjuntos de dados apresentado na Figura 22c não são diretamente comparáveis como ocorre para as discretizações anteriores. O perfil da distribuição no suporte entre as regras se assemelha à discretização comportamental e os valores de suporte máximo são próximos ao atingido na discretização quartis (Tabela 2). Contudo, há maior percentual de regras com suporte entre 0 e 1 que o verificado na discretização comportamental.

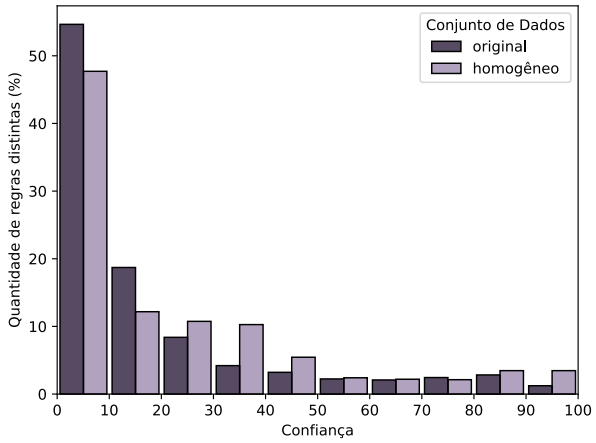
Figura 23 – Distribuição da confiança dividida entre os métodos de discretização comportamental, quartis e SAX para os conjuntos de dados original e homogêneo.

(a) Comportamental.

(b) Quartis.



(c) SAX.



Fonte: Elaborada pela autora.

Os gráficos da distribuição percentual das regras nas faixas de confiança divididos por método de discretização são apresentados na Figura 23 para ambos os conjuntos de dados. A discretização comportamental (Figura 23a) possui a distribuição percentual das regras mais

uniforme entre as faixas de confiança, principalmente para valores de confiança mais altos. Além disso, trata-se da discretização com menor percentual de regras com confiança entre 0 e 10.

A Figura 23b apresenta as distribuições para a discretização quartis, com o comportamento similar para ambos os conjuntos de dados. Embora o número de regras com confiança entre 50 e 90 seja baixo, para o conjunto de dados original são 3,14% regras geradas nessa faixa e 3,75% regras geradas no conjunto homogêneo, com quantidades que estão na mesma ordem do total de regras geradas pela discretização comportamental.

As distribuições das regras na confiança para a discretização SAX apresentadas na Figura 23c assemelham-se ao verificado na distribuição na discretização comportamental. O número de regras com confiança acima de 50 possui uma tendência de quantidade constante até a confiança de 80, mas há um leve aumento na quantidade de regras com confiança entre 80 e 100 que possuem suporte baixo (Figura 21a).

Através da avaliação entre os conjuntos de dados, verificou-se que o conjunto original e o homogêneo possuem alta similaridade, com comportamento semelhante no volume de regras geradas, suporte máximo alcançado e distribuição percentual no suporte e na confiança. O conjunto de dados original possui aproximadamente 50% mais séries que o conjunto homogêneo, com observações e variáveis faltantes. A distribuição das regras no suporte e na confiança do conjunto homogêneo apresenta um maior volume para valores mais elevados, mas com baixa variação no percentual em relação ao conjunto original.

Respondendo a primeira questão de pesquisa com as análises realizadas, entre as discretizações avaliadas, a comportamental apresentou os maiores valores de suporte máximo, melhor distribuição do suporte e da confiança no percentual de regras distintas, gerando o menor dispersão de regras distintas independentemente do conjunto de dados utilizados. A discretização SAX, embora apresente bons resultados quanto ao volume de regras geradas, quanto à aplicação de medidas de corte e quanto aos valores das métricas de avaliação, não se mostra como a mais indicada, dado que compreensão dos padrões das regras são específicos às suas ocorrências, demandando uma interpretação mais individualizada.

Embora a discretização quartis seja similar à comportamental por representar comportamentos de crescimento e decréscimo entre as observações, a discretização comportamental apresentou resultados superiores, com maior suporte máximo e menor volume de regras geradas mesmo utilizando um conjunto de dados com séries incompletas. Os resultados sugerem que a discretização quartis seja utilizada após a definição de comportamentos desejados nas regras, buscando-se então as regras na discretização quartis com detalhes desse comportamento. Para o restante da análise de dados faltantes será utilizada a discretização comportamental sobre o conjunto de dados homogêneo.

4.4.2 Avaliação do Impacto de Observações Faltantes

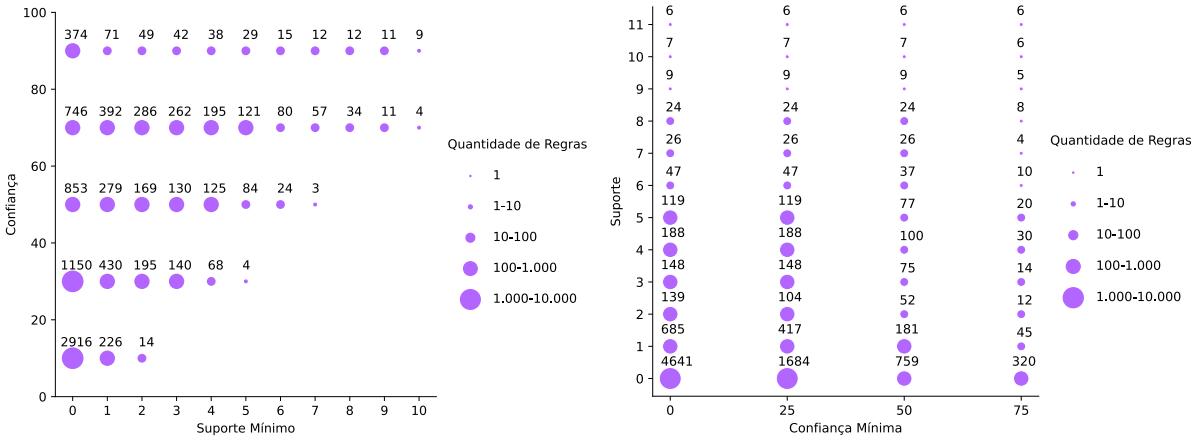
O eTRUMiner é capaz de lidar com séries multivariadas com observações e variáveis faltantes. Para avaliação do impacto de observações faltantes, utilizou-se o conjunto de dados homogêneo como referência e removeu-se observações de modo uniforme entre as variáveis e aleatoriamente dentro da variável. Os conjuntos de dados gerados possuem de 1% até 5%, 10% e 15% de observações faltantes. O objetivo nesta seção é avaliar a resposta do algoritmo frente aos dados faltantes e por isso definiu-se como método de discretização o comportamental por ter apresentado os melhores resultados nas análises anteriores.

As distribuições das regras para a discretização comportamental após a aplicação de corte no conjunto homogêneo é apresentado na Figura 24 para indicar a adequação do método sobre esse conjunto. Na Figura 24a, o corte no suporte $sup_{min} = 8$ retorna 46 regras, com a principal faixa de confiança, contendo $73,91\%$ das regras remanescentes, também de 60 a 80. Contudo, para $sup_{min} = 10$ no conjunto homogêneo a principal faixa de confiança é de 80 a 100, com 13 regras remanescentes, demonstrando que as regras mais resilientes aos cortes apresentam a confiança na faixa mais elevada.

Figura 24 – Distribuição das regras após aplicação de corte na **discretização comportamental** para o conjunto de dados homogêneo.

(a) Corte no suporte.

(b) Corte na confiança.



Fonte: Elaborada pela autora.

Verifica-se na Figura 24b que todas as regras na maior faixa de suporte 11 não são eliminadas por nenhum corte na confiança, enquanto regras com suporte 10 diminuem em apenas 15% para $conf_{min} = 75$. As regras produzidas pelo conjunto homogêneo com discretização comportamental possuem maior resiliência a cortes no suporte e na confiança, indicando que tratam-se de regras fortes e que o método é adequado para o conjunto.

A Tabela 3 contém o volume de regras distintas geradas pelo eTRUMiner entre o conjunto de dados homogêneo completo e o mesmo conjunto de dados com remoções variando de 1%

a 15%. Embora para percentuais pequenos não se verifique grande perda no volume, para a remoção de 15% há uma redução de 20% na quantidade de regras.

Tabela 3 – Número de regras distintas por percentual de remoção do conjunto de dados homogêneo de 1% a 5%, 10% e 15% de observações removidas.

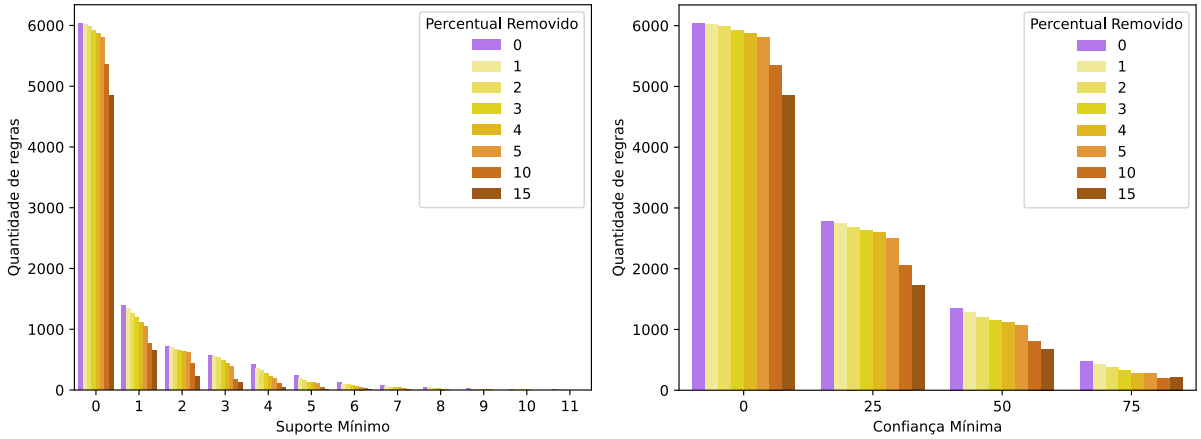
Percentual Removido	**Número de Regras**
Homogêneo	6.039
1%	6.017
2%	5.986
3%	5.924
4%	5.867
5%	5.806
10%	5.351
15%	4.852

Fonte: Dados da pesquisa.

A Figura 25 apresenta o número de regras para cada corte no suporte e na confiança nos percentuais de remoção do conjunto de dados homogêneo. O maior suporte mínimo avaliado foi $sup_{min} = 11$ que retorna 6 regras no conjunto de dados homogêneo completo, e ainda há retorno de regras para os conjuntos com a remoção de até 2%. Observa-se na Figura 25a que a remoção de observações produz em média regras com suportes menores, de modo que os cortes no suporte devem ser mais baixos para que não ocorra a remoção total das regras. O corte $sup_{min} = 1$ já reduz a quantidade de regras para percentuais entre $23,15\%$ e $13,38\%$ da quantidade original.

Figura 25 – Número de regras do conjunto de dados homogêneo com e sem percentuais de observações removidas para cortes nas métricas de avaliação.

(a) Suporte Mínimo. (b) Confiança Mínima.



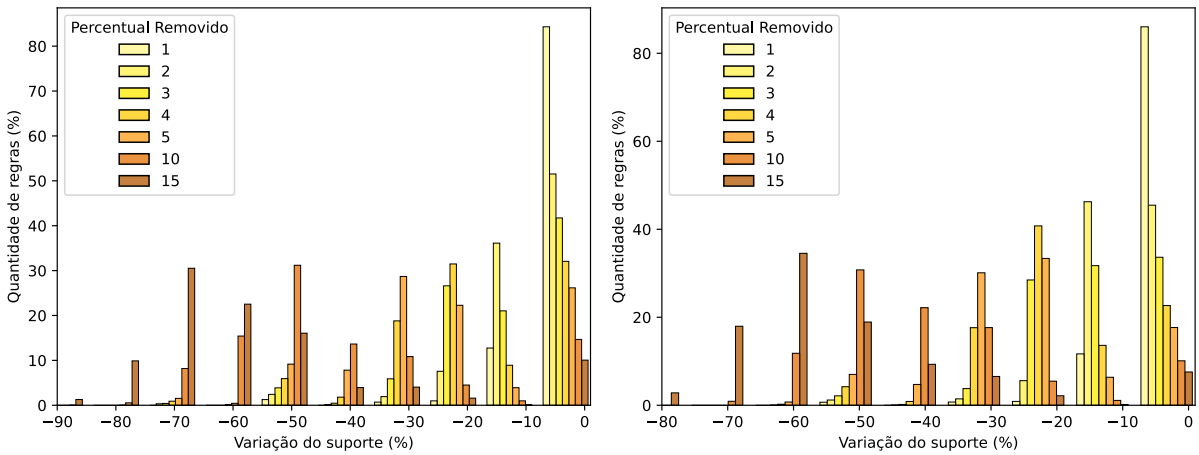
Fonte: Elaborada pela autora.

Os cortes na confiança são apresentados na Figura 25b variando de 0 a 75. A aplicação do corte nessa métrica reduz aproximadamente pela metade o volume de regras geradas para

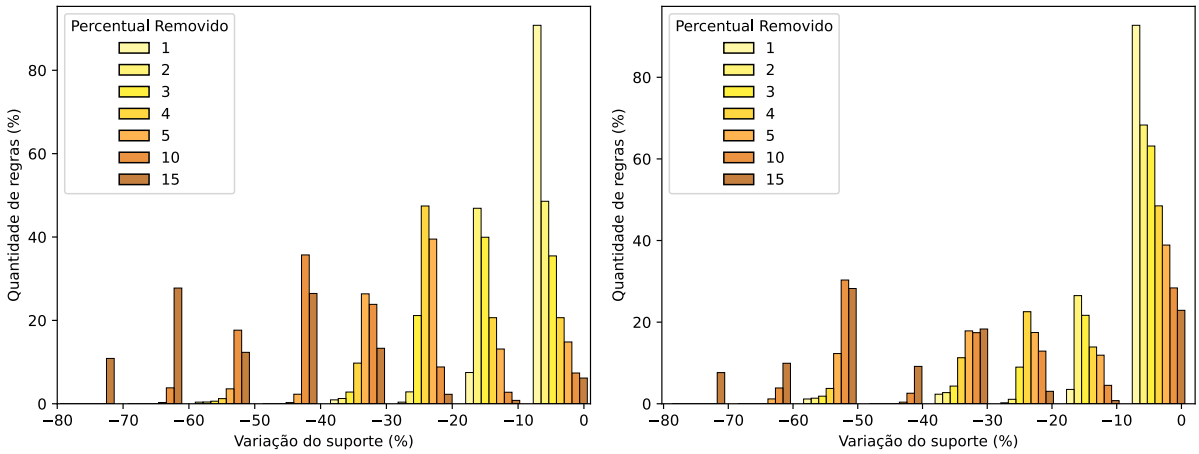
$conf_{min}$ = 25 independentemente do percentual de remoção avaliado. Além disso, dentro de um mesmo corte, a quantidade de regras entre os percentuais de remoção mantém a proporção existente, exceto para $conf_{min}$ = 75. Em uma comparação entre o corte no suporte e na confiança, o corte na confiança mostra-se menos punitivo aos conjuntos de dados com remoção.

Figura 26 – Variação percentual do suporte entre conjunto de dados com observações removidas de 1% a 15% com corte na confiança e $sup_{min} = 0$.

(a) $conf_{min} = 0$. (b) $conf_{min} = 25$.



(c) $conf_{min} = 50$. (d) $conf_{min}$= 75.



Fonte: Elaborada pela autora.

Para compreender melhor como as regras obtidas no conjunto de dados homogêneo completo se alteram com a presença de observações faltantes, avaliou-se a variação percentual do suporte para as regras coincidentes em cada conjunto de dados com percentual removido. A Figura 26 contém os resultados para corte na confiança variando de 0 a 75 mas sem corte no suporte.

A Figura 26a apresenta a distribuição da variação percentual entre o conjunto de dados homogêneo completo e os conjuntos com percentuais removidos para as regras sem nenhum corte. O número de regras coincidentes varia de 6.017 a 4.852 para os percentuais de 1% a

15%, que é exatamente o total de regras obtido em cada conjunto com remoção, apresentado na Tabela 3. Ou seja, não há geração de novas regras para os conjuntos com dados faltantes usando a discretização comportamental. As médias da variação percentual do suporte são de $-4,24\%$ até $-55,17\%$ no conjunto de dados com 15% de observações faltantes, conforme Tabela 4, indicando que no conjunto com 15% de remoção, as regras remanescentes obtidas deste conjunto possuem em média um suporte 55% inferior ao valor obtido no conjunto sem dados faltantes.

Tabela 4 – Média da variação percentual do suporte para regras coincidentes entre conjuntos de dados com remoção de observações para corte na confiança.

Percentual Removido	**conf$_{min}$ = 0**	**conf$_{min}$ = 25**	**conf$_{min}$ = 50**	**conf$_{min}$ = 75**
1%	-4,24%	-4,45%	-4,17%	-3,51%
2%	-8,60%	-8,72%	-7,99%	-5,52%
3%	-13,03%	-13,08%	-11,79%	-7,56%
4%	-18,66%	-19,46%	-18,23%	-12,78%
5%	-24,03%	-24,68%	-23,67%	-19,00%
10%	-41,60%	-40,67%	-39,27%	-27,97%
15%	-55,17%	-52,71%	-50,63%	-37,26%

Fonte: Dados da pesquisa.

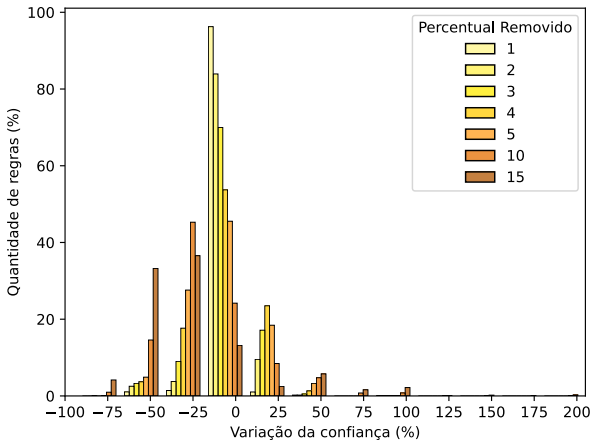
Para o conjunto com 1% de dados removidos, mais de 90% das regras variam apenas 10% do suporte em relação o conjunto de dados homogêneo completo. Com 15% de observações faltantes, $4,26\%$ das regras coincidentes apresentam mais de 80% de variação do suporte, sendo que a regra com maior variação ($94,36\%$) compõe-se de todas as variáveis e possui suporte original de $sup = 0,25$. O conjunto de dados com percentual de observações removidas semelhante ao conjunto de dados original (10% de remoção) varia o suporte (entre as regras coincidentes) em média de 41,60%, verificando-se um suporte até 85,64% menor que o suporte do conjunto de dados homogêneo completo.

Para cortes menores, verifica-se um aumento no percentual de regras com maior variação do suporte em conjuntos de dados com baixos índices de dados faltantes, refletindo na média de variação percentual. Contudo, a partir de $conf_{min} = 50$, as médias de todos os conjuntos diminuem, atingindo $-37,26\%$ de variação percentual do suporte para $conf_{min} = 75$ no conjunto de dados com 15% de observações removidas. Esse perfil indica que embora o corte na confiança não reduza drasticamente o volume de regras, sua aplicação é capaz de selecionar as regras com menor variação no suporte em relação ao suporte do conjunto completo, principalmente para conjuntos com grande quantidade de observações faltantes.

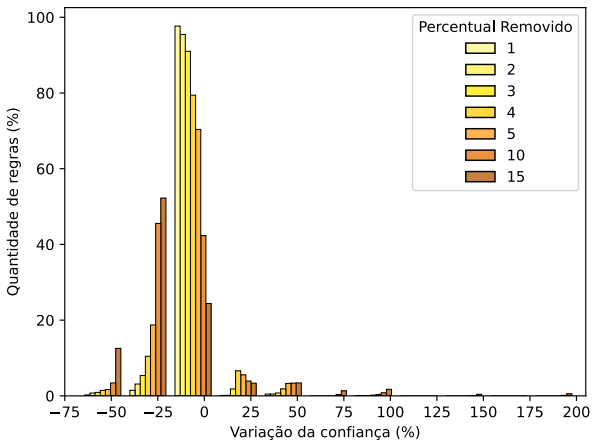
Para a análise da variação percentual da confiança entre regras coincidentes do conjunto de dados homogêneo completo e com remoção de observações, avaliou-se os resultados retornados sem corte no suporte e com corte na confiança de 0 a 75. As distribuições são apresentadas na Figura 27, com $conf_{min} = 0$ na Figura 27a, $conf_{min} = 25$ na Figura 27b, $conf_{min} = 50$ na Figura 27c e a Figura 27d para $conf_{min} = 75$.

Figura 27 – Variação percentual da confiança entre conjunto de dados com observações removidas de 1% a 15% com corte na confiança e $sup_{min} = 0$.

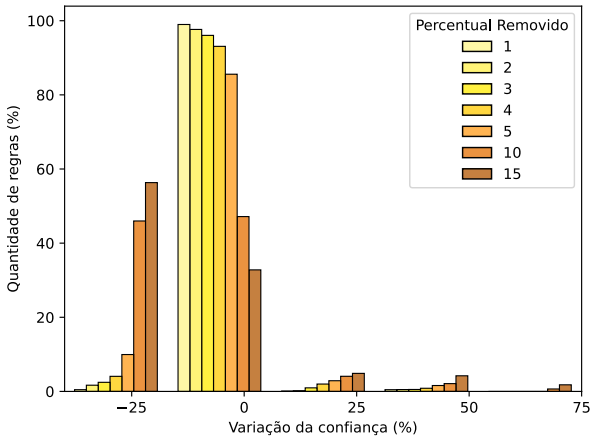
(a) $conf_{min} = 0$.



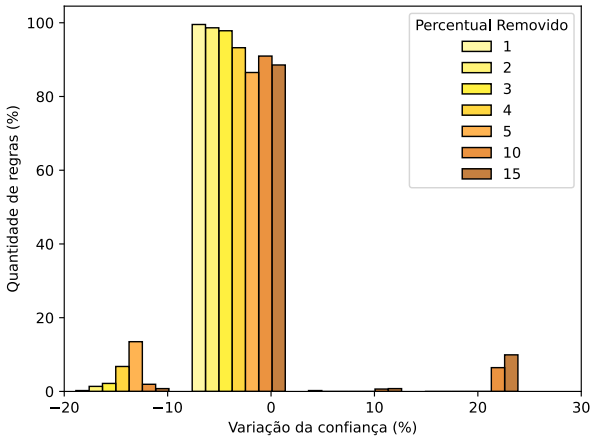
(b) $conf_{min} = 25$.



(c) $conf_{min} = 50$.



(d) $conf_{min} = 75$.



Fonte: Elaborada pela autora.

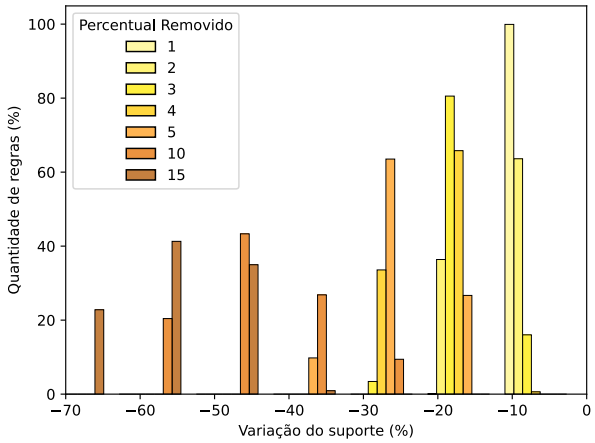
O primeiro corte na confiança dispersa a distribuição da variação percentual da confiança em conjuntos de dados com maiores percentuais de remoção, com mais metade das regras com a confiança percentual variando de -50% a -25% entre regras coincidentes no conjunto de dados com 15% de remoção. Com o corte $conf_{min} = 50$, a diferença entre os valores mínimo e máximo da média da variação da confiança é de apenas $10,73$.

No maior corte na confiança, $conf_{min} = 75$ apresentado na Figura 27d, há baixa variabilidade na variação da confiança, como verificado na última coluna da Tabela 5. Contudo, o número de regras remanescentes máximo é de 427 para o conjunto de dados com 1% de remoção, que representa $7,10\%$ das regras retornadas no conjunto sem observações faltantes. Das 131 regras retornadas pelo conjunto de dados com 15% de remoção e corte de $conf_{min} = 75$, aproximadamente 90% apresenta variação percentual da confiança entre -5% e 5%.

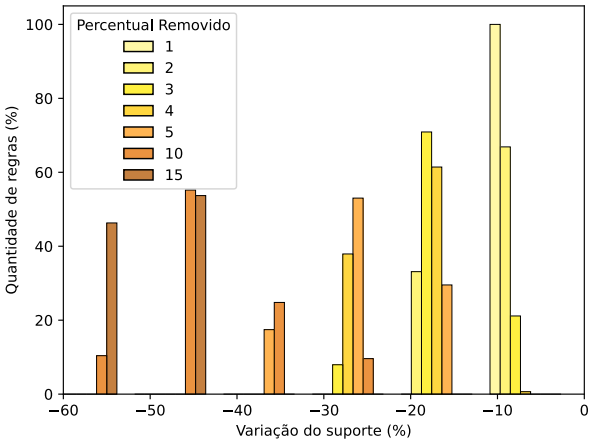
Assim como verificado na variação percentual do suporte, o corte na confiança seleciona as regras mais semelhantes (quanto aos valores das métricas de avaliação) às obtidas no conjunto homogêneo completo. Contudo, a redução no volume não é suficiente, e por isso, deve ser aliada ao corte no suporte.

Figura 28 – Variação percentual do suporte entre conjunto de dados com observações removidas de 1% a 15% com corte no suporte e na confiança.

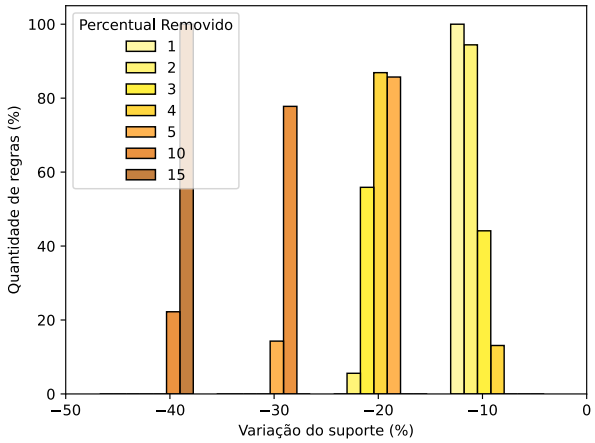
(a) $sup_{min} = 1$, $conf_{min} = 0$.



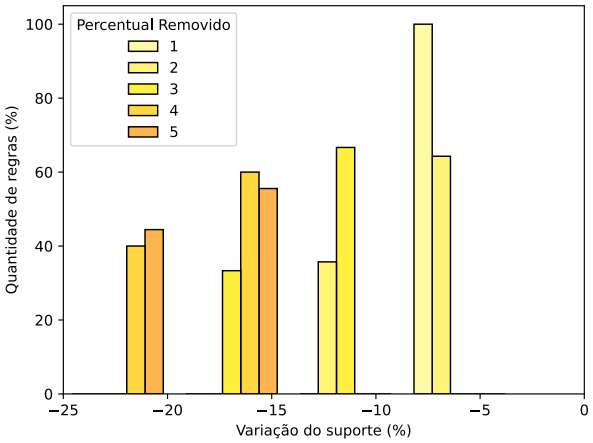
(b) $sup_{min} = 2$, $conf_{min} = 25$.



(c) $sup_{min} = 4$, $conf_{min} = 50$.



(d) $sup_{min} = 6$, $conf_{min} = 75$.



Fonte: Elaborada pela autora.

Na Figura 28 apresenta-se diferentes cortes no suporte e na confiança para avaliar o comportamento da variação do percentual do suporte entre as regras coincidentes sobre cortes múltiplos. Todos os cortes avaliados possuem 100% das regras retornadas pelos conjuntos de dados com remoção coincidentes com o conjunto de dados homogêneo completo, sendo as regras remanescentes aquelas que possuem menor variação percentual do suporte. Contudo, o aumento nos cortes reduz drasticamente a quantidade de regras comuns.

A Figura 28a apresenta a variação percentual do suporte na faixa de 0 a -70 (para o menor corte avaliado, $sup_{min} = 1$, $conf_{min} = 0$), enquanto o maior corte avaliado, $sup_{min} = 6$, $conf_{min} = 75$ (Figura 28d) varia de -5 a -25 apenas. Portanto, o aumento nos cortes retorna as regras com mínima variação no suporte e após análise, verifica-se que são regras com maior suporte e alta confiança no conjunto completo. O corte no suporte varia entre 1 e 6, que é o maior valor que ainda retorna regras do conjunto de dados com maior percentual de remoção quando $conf_{min} = 0$.

O corte $sup_{min} = 1$, $conf_{min} = 0$ retorna 1.345 regras do conjunto de dados com 1% de observações removidas e 649 regras para 15% de remoção, enquanto o corte $sup_{min} = 6$, $conf_{min} = 75$ sobre o conjunto de dados com menor percentual de remoção retorna apenas 26 regras do total de 6.017. Na Figura 28d, observa-se que não há regras coincidentes para os conjuntos de dados com 10% e 15% de remoção, já que para esses cortes não há retorno de nenhuma regra pelo algoritmo.

Tabela 5 – Média da variação percentual da confiança para regras coincidentes nos conjuntos de dados com remoção de 1% a 15% de observações para cortes na confiança.

Percentual Removido	**conf$_{min}$ = 0**	**conf$_{min}$ = 25**	**conf$_{min}$ = 50**	**conf$_{min}$ = 75**
1%	-1,71%	-1,87%	-1,32%	-0,67%
2%	-3,81%	-4,09%	-2,96%	-1,33%
3%	-5,73%	-5,86%	-4,29%	-1,66%
4%	-7,56%	-7,50%	-6,27%	-2,20%
5%	-9,74%	-8,41%	-6,90%	-2,19%
10%	-18,90%	-14,20%	-11,12%	0,95%
15%	-25,06%	-15,25%	-12,05%	2,25%

Fonte: Dados da pesquisa.

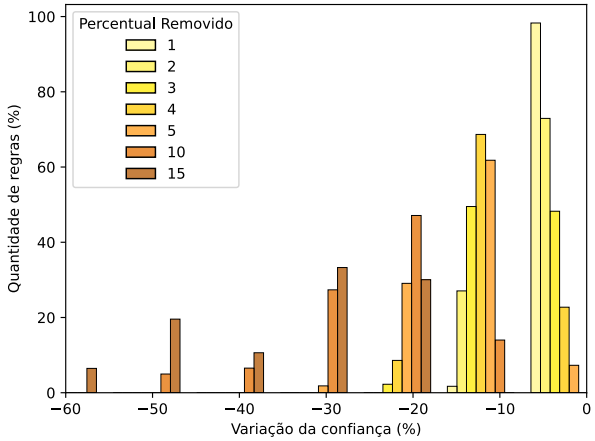
A Tabela 5 apresenta os valores médios da variação percentual da confiança para cortes na confiança entre conjuntos de dados com 1% até 15% de observações removidas. A média da variação percentual da confiança abrange de $-1,71\%$ até $-25,06\%$ para os conjuntos de dados com 1% removido até 15% de remoção, respectivamente (para as regras sem corte).

A Figura 29 apresenta a variação percentual da confiança para regras retornadas pelos conjuntos de dados com observações removidas após corte no suporte e na confiança. Os cortes aplicados na confiança são os mesmos apresentados na Figura 27. A aplicação de corte no suporte mostra-se como mais efetiva na remoção de regras com alta variabilidade na confiança, como observa-se comparando a Figura 27b (o menor corte na confiança) abrangendo a variação percentual de -75% até 200%, e a Figura 29a com corte apenas no suporte, em que a confiança varia entre -60% e 0%.

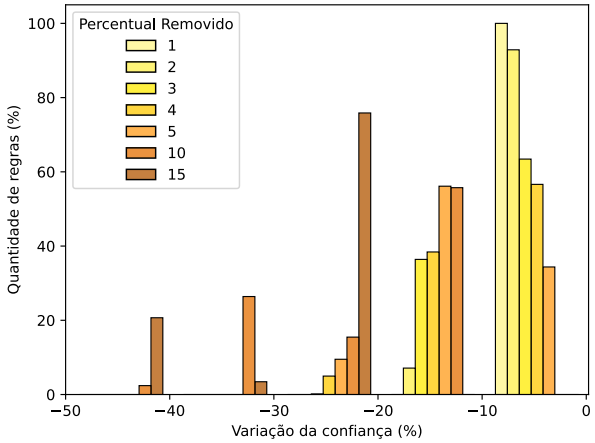
Para o maior corte, $sup_{min} = 6$, $conf_{min} = 75\%$ não são retornadas regras pelos dois conjuntos de dados com maior percentual de remoção, como já verificado na Figura 28d, mas a variabilidade percentual da confiança entre regras coincidentes dos conjuntos de dados

Figura 29 – Variação percentual da confiança entre conjuntos de dados com observações removidas de 1% a 15% com corte no suporte e na confiança.

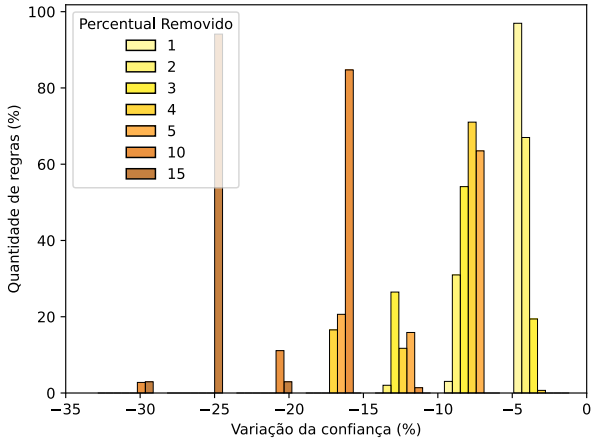
(a) $sup_{min} = 1$, $conf_{min} = 0$.



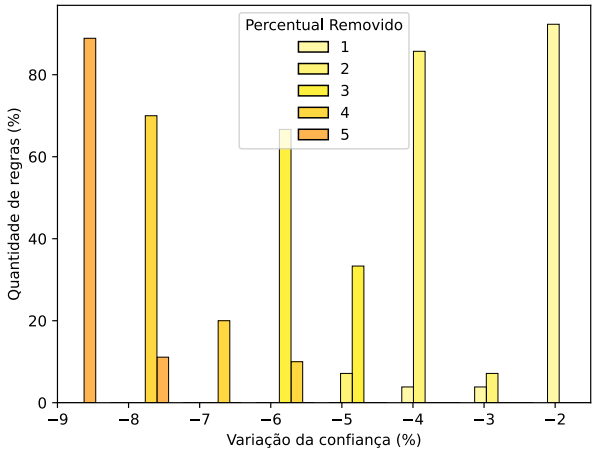
(b) $sup_{min} = 2$, $conf_{min} = 25$.



(c) $sup_{min} = 4$, $conf_{min} = 50$.



(d) $sup_{min} = 6$, $conf_{min} = 75$.



Fonte: Elaborada pela autora.

remanescentes é muito baixa, mostrando que as regras retornadas possuem métricas de avaliação muito próximas às obtidas no conjunto completo.

Dentre os cortes avaliados, o mais alto que retorna regras dos conjuntos de dados com 10% e 15% de remoção é $sup_{min} = 4$, $conf_{min} = 50$ com 72 e 34 regras respectivamente. Embora represente uma quantidade pequena do total de regras geradas pelos conjuntos, é um volume alto de regras coincidentes com o conjunto de dados homogêneo completo que possuem baixa variação de suporte e confiança e também apresentam suporte e confiança elevados.

Conjuntos de dados com observações faltantes retornam resultados muito semelhantes ao conjunto completo, com 100% de intersecção entre regras temporais obtidas e alta semelhança nos valores de suporte e confiança entre regras coincidentes. O uso de corte nas métricas de

avaliação permite selecionar de forma eficiente as regras mais semelhantes às regras do conjunto completo, com baixa penalização devido ao uso de um conjunto incompleto. Por tanto, trabalhar com conjuntos de dados com observações faltantes é promissor e verifica-se que o eTRUMiner é capaz de obter as regras de interesse (alto suporte e alta confiança) mesmo nesses cenários.

O conjunto de dados original, como descrito na Seção 4.1, possui $9,39\%$ de observações faltantes e, com base na tendência dos dados verificado para o conjunto de dados homogêneo, estima-se que ele gere 10% a menos de regras distintas em relação ao que geraria o conjunto original completo. Dado a baixa perda percentual, há grande interesse em analisar o conjunto de dados original mesmo com as observações faltantes, justificando o desenvolvimento e uso do eTRUMiner.

Respondendo a segunda questão de pesquisa, a discretização comportamental permite a geração de regras mais similares às geradas pelo conjunto completo pois o método não altera a série discretizada quando há a presença de observações faltantes. Nas análises utilizando conjuntos com dados removidos, verifica-se que há baixa penalização do uso de um conjunto incompleto, com a aplicação de cortes mostrando-se como uma alternativa para selecionar as regras de interesse coincidentes. A Subseção 4.4.1 indica que mesmo com o conjunto heterogêneo, o impacto nas regras é aceitável.

4.5 Análise Semântica

O último conjunto de experimentos visa analisar semanticamente o conjunto de dados do comércio internacional. As principais regras entre as discretizações comportamental e quartis, que são métodos semelhantes quanto ao significado das regras, e a característica temporal mais comum entre as regras temporais são discutidas na Subseção 4.5.1. Na Subseção 4.5.2, divide-se o intervalo das séries temporais do conjunto homogêneo em 3 períodos de interesse, e avalia-se com maiores detalhes as regras geradas pela discretização comportamental.

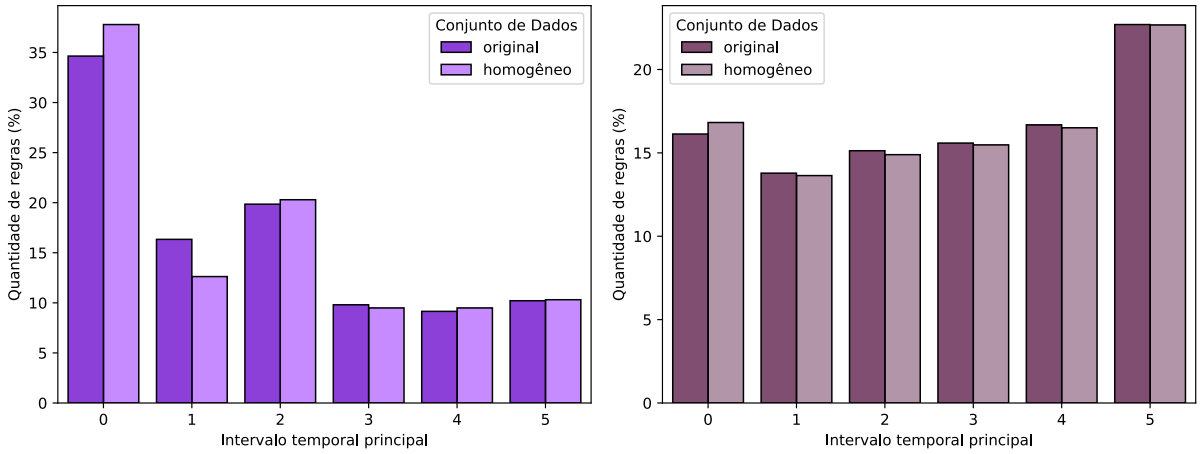
4.5.1 Avaliação de Regras Genéricas

As regras genéricas são regras temporais agrupadas pelo antecedente e consequente comuns, avaliadas para melhor compreensão da regras sem influência da dispersão gerada pela característica temporal (Δt). Cada regra genérica é gerada por até $w+1$ regras temporais, cada qual com seu suporte e confiança. O intervalo temporal principal de uma regra genérica é determinado pela característica temporal da regra que apresenta maior valor de suporte e confiança dentre as regras temporais que constituem a regra genérica. Por exemplo, a regra genérica $[IMP,I] \rightarrow [PIB,I]$ possui como intervalo temporal principal $\Delta t = 0$, dado que a regra $([IMP,I] \rightarrow [PIB,I], \Delta t = 0)$ possui $sup = 9,07$ e $conf = 74,53$, que é o maior valor de suporte e confiança entre as regras de aumento na importação seguido por aumento no PIB para qualquer característica temporal.

Para a análise dos intervalos temporais principais mais frequentes entre as regras, verificou-se a distribuição percentual para cada conjunto de dados e método de discretização, conforme Figura 30. Entre os conjuntos de dados não há grande diferença na distribuição, independentemente do método de discretização utilizado.

Figura 30 – Distribuição percentual do intervalo temporal principal entre regras genéricas.

(a) Comportamental. (b) Quartis.



Fonte: Elaborada pela autora.

Quanto à distribuição, o método quartis tende a uma distribuição quase constante entre intervalos temporais principais, com aumento no maior valor, indicando uma homogeneidade na frequência de cada intervalo temporal principal das regras genéricas, com pequena predominância do maior Δt. O método comportamental apresenta um perfil decrescente com aumento da característica temporal, indicando predominância das regras com intervalo temporal nulo.

A Figura 30a apresenta a distribuição das regras genéricas no intervalo temporal principal para a discretização comportamental utilizando os conjuntos de dados original e homogêneo. Cerca de 35% das regras genéricas contêm a regra temporal com $\Delta t = 0$ como a regra com maior suporte e confiança do conjunto, indicando $\Delta t = 0$ como intervalo temporal principal de ambos os conjuntos de dados. Para intervalos de 3 até 5 anos, o percentual reduz a 10% das regras.

Para o método quartis, Figura 30b, o aumento ocorre em $\Delta t = 5$, abrangendo 20% do total das regras genéricas. Essa predominância dos intervalos temporais principais com intervalos maiores é contrário ao comportamento esperado já que são as regras menos geradas pelo algoritmo, e indicam que ou as regras temporais com característica temporal mais altas apresentam os maiores valores de suporte e confiança ou há alto volume de regras temporais com característica temporal majoritariamente alta.

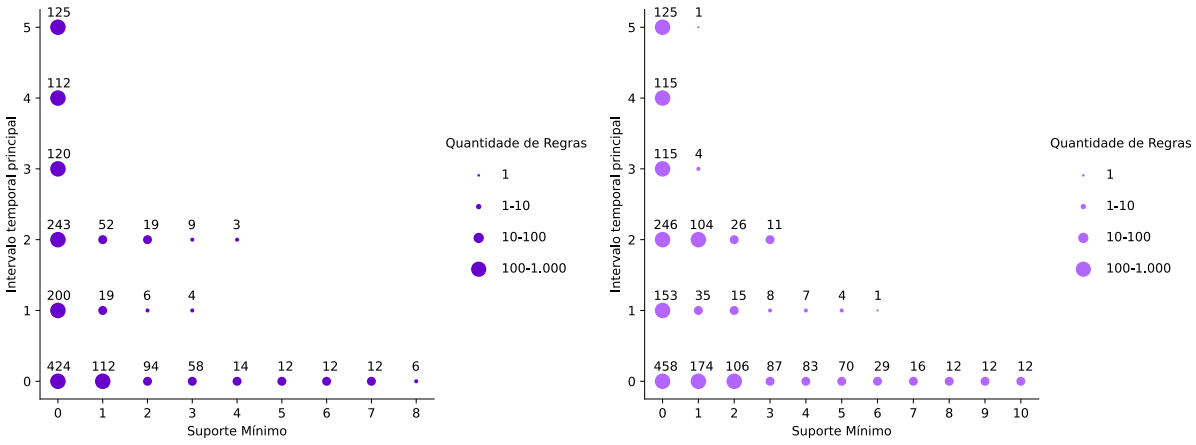
O perfil da distribuição do intervalo temporal principal entre as regras genéricas esperado é de um pico em $\Delta t = 0$ e queda proporcional ao número de séries e variáveis com o aumento do

intervalo temporal. Dentre as discretizações avaliadas, a comportamental é a que mais coincide com o esperado, mesmo sem a aplicação de corte nas métricas de avaliação.

Para analisar o intervalo temporal principal das regras relevantes, foram avaliados vários cortes no suporte das regras temporais, em consonância com a distribuição do suporte específica para cada discretização e conjunto de dados utilizados. Verificou-se as regras genéricas para cada corte no suporte e o intervalo temporal principal dentro do conjunto de regras temporais retornadas. A Figura 31 e a Figura 32 apresentam os resultados obtidos para as discretizações comportamental e quartis.

Figura 31 – Distribuição do intervalo temporal principal para regras genéricas na discretização comportamental com corte no suporte.

(a) Conjunto original. (b) Conjunto homogêneo.



Fonte: Elaborada pela autora.

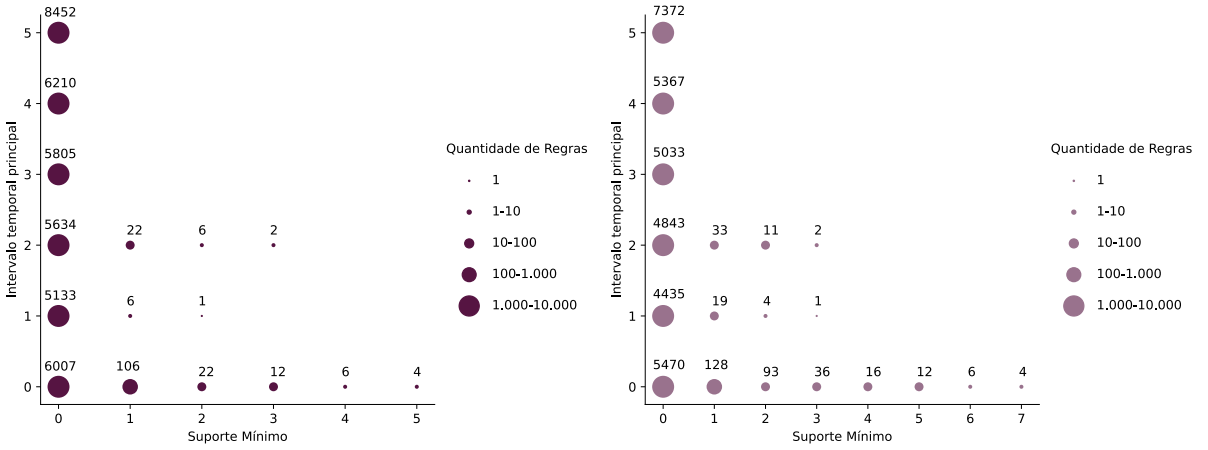
Na discretização comportamental, o intervalo temporal principal mantém-se $\Delta t = 0$ independentemente do conjunto de dados e do suporte mínimo definido. Com o aumento no corte, a quantidade total de regras genéricas cai de 1.224 para 6 para o conjunto original (Figura 31a) e de 1.212 para 12 no conjunto homogêneo (Figura 31b). A partir de $sup_{min} = 5$ no conjunto de dados original e $sup_{min} = 7$ no conjunto homogêneo, o único intervalo temporal principal é $\Delta t = 0$.

A principal regra genérica para o conjunto de dados original é "com o aumento na importação, verifica-se o aumento no PIB", ocorrendo majoritariamente no mesmo ano com suporte de $15,77$ e confiança de $67,72$. A principal regra genérica se mantém para o conjunto de dados homogêneo e o suporte sobe para $58,45$, com a confiança atingindo $79,68$.

O intervalo temporal principal predominante na discretização quartis é o mais alto, $\Delta t = 5$. Para ambos os conjuntos de dados, o primeiro corte no suporte avaliado, $sup_{min} = 1$, remove essa tendência, com predominância para $\Delta t = 0$ como intervalo principal. Além disso, todas as regras genéricas com intervalo temporal principal maior que 3 são eliminadas.

Figura 32 – Distribuição do intervalo temporal principal para regras genéricas na discretização quartis com corte no suporte.

(a) Conjunto original. (b) Conjunto homogêneo.



Fonte: Elaborada pela autora.

A principal regra genérica da discretização quartis em ambos os conjuntos de dados é “um aumento de até 25% na importação é seguido por um aumento de até 25% no PIB”, ocorrendo principalmente no mesmo ano. No conjunto de dados original o suporte é de $29,95$ e a confiança de $64,62$, enquanto no homogêneo $sup = 39,49$ e $conf = 72,93$. Essa regra coincide com a regra genérica verificada na discretização comportamental, indicando que grande parte da regra $[IMP,I] \Rightarrow [PIB,I]$ é de aumento de até 25% tanto no antecedente quanto no consequente.

Ao analisar as regras genéricas, verifica-se valores de suporte mais elevados, mostrando que a principal influência para a obtenção de baixo suporte é a característica temporal presente nas regras temporais, que aumenta significativamente o número de transações, conforme detalhado na Seção 3.1. Na análise de regras genéricas, a discretização comportamental se mostra como a mais adequada, sem a necessidade da aplicação de cortes, com um número coerente de regras e métricas de avaliação promissoras. A discretização quartis apresenta-se como auxiliar à mineração de regras com a aplicação da discretização comportamental: as regras encontradas ao utilizar o método comportamental podem ser mais detalhadas pelas regras obtidas pelo método quartis.

4.5.2 Avaliação de Períodos de Interesse

Para realizar a análise semântica, utilizou-se o conjunto de dados homogêneo. O conjunto foi dividido em 3 subconjuntos de períodos consecutivos e distintos: 1996 a 2004, 2005 a 2012 e 2013 a 2019, abrangendo respectivamente a crise da bolha de internet[6] e sua recuperação

[6] Dot-Com Bubble <https://www.goldmansachs.com/our-firm/history/moments/2000-dot-com-bubble.html>

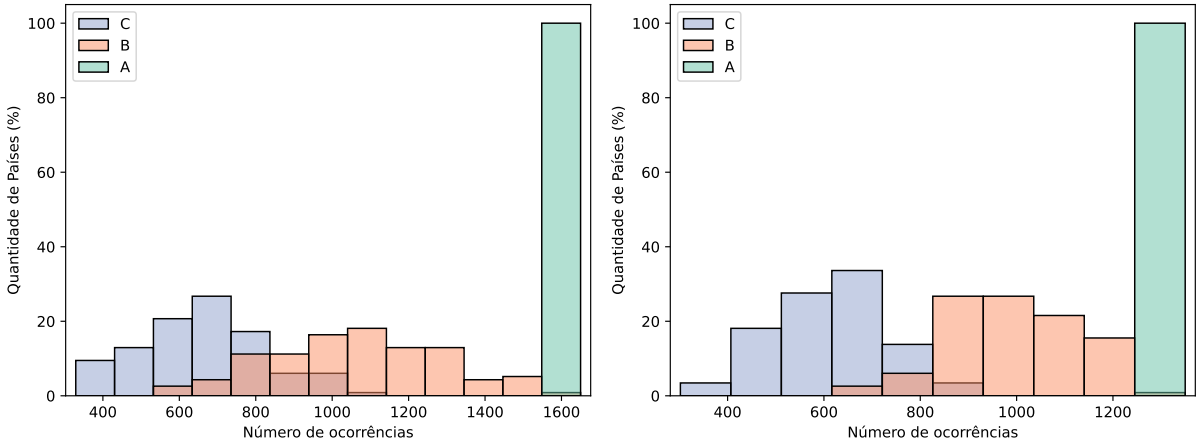
econômica, a Grande Recessão Econômica de 2009[7] e um período de instabilidade econômica pré-covid (UNCTAD, 2023).

A análise adequada de cada período requer a aplicação de um corte nas regras por meio das métricas de avaliação capaz de reduzir significativamente o número de regras retornadas e selecionar as regras mais relevantes. Conforme Figura 24a, no conjunto de dados homogêneo com discretização comportamental, o primeiro suporte mínimo que elimina as regras com confiança abaixo de 20 mas mantém grande parte das outras regras é $sup_{min} = 3$.

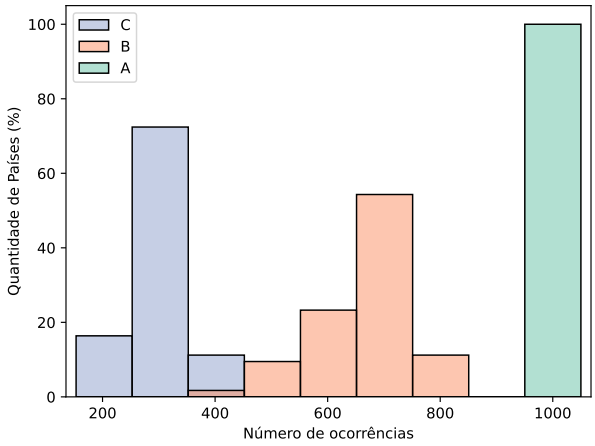
Figura 33 – Número de ocorrências por percentual de países para cortes no suporte e na confiança por período.

(a) 1996 a 2004.

(b) 2005 a 2012.



(c) 2013 a 2019.



Fonte: Elaborada pela autora.

A Figura 33 apresenta a avaliação do corte $sup_{min} = 3$, $conf_{min} = 50$ indicado por "C" e um corte menos drástico de $sup_{min} = 2$, $conf_{min} = 25$, indicado por "B". Os resultados do conjunto sem corte está representado por "A". Os gráficos apresentam, para cada corte, a

[7] Great Recession <https://www.federalreservehistory.org/essays/great-recession-and-its-aftermath>

distribuição do número de ocorrências das regras geradas pelo eTRUMiner pela porcentagem de países. Observa-se que, no cenário sem corte, as regras extensas retornadas pelo algoritmo contém 1600 ocorrências em todos os países, no período de 1996 a 2004. Com o corte de $sup_{min} = 2$, $conf_{min} = 25$, há 1% dos países apresentando um máximo de 1600 ocorrências e 3% dos países com um mínimo de 600 ocorrências, indicando um corte significativo no número de regras. Nota-se, numa análise similar, que um corte mais acentuado no suporte poderia acarretar em países sem regras geradas, já que no período de 2013 a 2019 a redução chega a 80% para $sup_{min} = 3$, $conf_{min} = 50$ conforme Figura 33c. Contudo, o corte mais brando avaliado, ainda retém um elevado número de ocorrências para alguns países, e portanto, o corte definido para as análises nesta seção é $sup_{min} = 3$, $conf_{min} = 50$.

O volume de regras antes e após a aplicação do corte baseado nas métricas de avaliação e o suporte máximo observado é apresentado na Tabela 6. O suporte máximo verificado entre os períodos é de $18,84$ obtido de 2005 a 2012, e o número de regras distintas sem corte retornadas por período varia entre 3.409 e 4.713. O período de 2013 a 2019 mostra-se como o maior gerador de regras distintas e o menor suporte máximo, indicando tratar-se de uma época com comportamento econômico mais difuso entre os países. Já de 2005 a 2012 nota-se maior tendência mundial nos dados, com menor variabilidade de regras e maior valor de suporte máximo.

Tabela 6 – Número de regras distintas com e sem a aplicação de corte nas métricas de avaliação e suporte máximo entre períodos.

	Sem Corte	**Com Corte**	**Suporte Máximo**
1996 a 2004	4.160	340	15,54
2005 a 2012	3.409	345	18,84
2013 a 2019	4.713	264	13,38

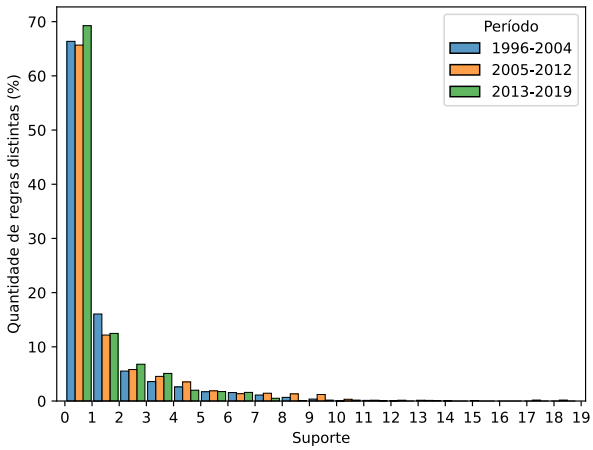
Fonte: Dados da pesquisa.

A Figura 34 apresenta a distribuição do suporte e da confiança entre os períodos, antes e após o corte de $sup_{min} = 3$, $conf_{min} = 50$. A redução drástica observada na Tabela 6 para os resultados com a aplicação do corte é devido ao grande volume de regras com suporte baixo, entre 0 e 1 na Figura 34a, e confiança reduzida, abaixo de 10 na Figura 34c.

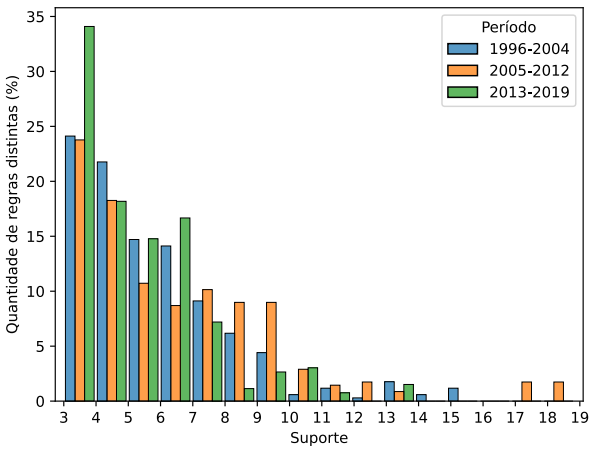
O primeiro período (1996-2004) apresenta um pico no suporte abaixo de 1 (Figura 34a), seguido de uma queda uniforme cujo perfil de queda se mantém mesmo após a aplicação do corte (Figura 34b). O período intermediário (2005-2012) que abrange a Grande Recessão Econômica apresenta dois picos anormais na distribuição do suporte, entre $7-10$ e $17-19$, indicando uma maior densidade de regras nesses valores. Verifica-se uma queda mais suave nos suportes mais altos das regras geradas, sugerindo a presença de maior volume de regras mais fortes que nos demais períodos. Ao analisar a Figura 34b, nota-se que o período de 2013 a 2019 apresenta predominância de regras na faixa de suporte entre $3-4$ após a aplicação de corte no suporte e na confiança, indicando que além do comportamento mais difuso que acaba por gerar um maior volume de regras distintas, o suporte do período também se concentra nas faixas mais baixas.

Figura 34 – Distribuição das regras entre métricas de avaliação nos períodos sem corte e $sup_{min} = 3$, $conf_{min} = 50$.

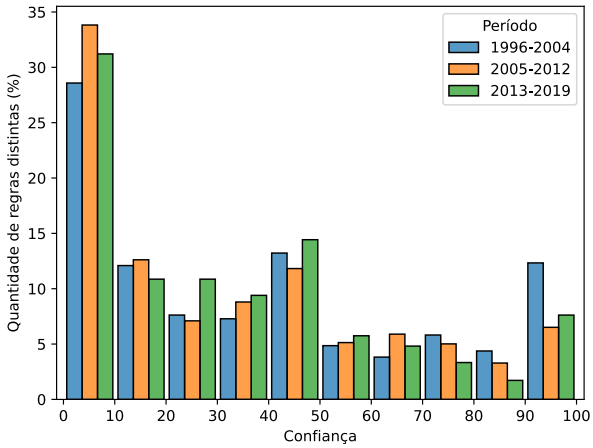
(a) Distribuição do suporte sem corte.



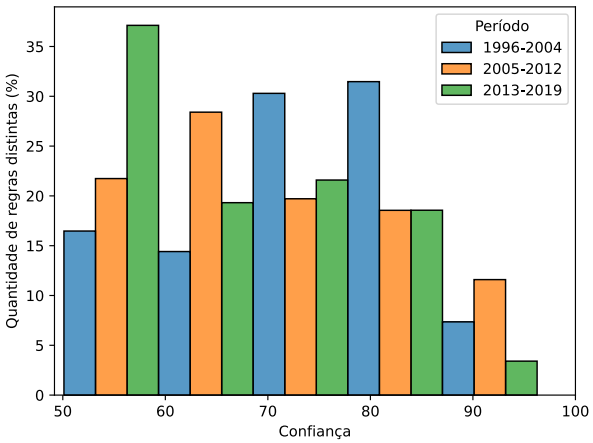
(b) Distribuição do suporte com corte.



(c) Distribuição da confiança sem corte.



(d) Distribuição da confiança com corte.



Fonte: Elaborada pela autora.

A distribuição das regras na confiança por período com a aplicação do corte no suporte e na confiança é apresentado na Figura 34d. O primeiro período possui a faixa predominante de confiança mais alta, entre 70 e 90, apontando que as regras retornadas apresentam maior certeza quanto ao consequente em relação aos demais períodos. O período de 2005 a 2012 possui um pico de regras entre 60 e 70, indicando maior incerteza quanto ao consequente entre as regras. No intervalo de 2013 a 2019 o pico de regras é na faixa de confiança entre 50 e 60, mesmo após o corte, destacando a instabilidade econômica existente no período.

Para compreender o comportamento dos países por período, definiu-se uma análise quantitativa, que utiliza as regras extensas contendo a localização nas séries que ocorrem. O conjunto de regras com corte $sup_{min} = 3$, $conf_{min} = 50$ foi dividido em 4 quartis para avaliação da distribuição percentual das ocorrências de cada país entre esses quartis. O 1° quartil abrange

as regras com maior suporte e maior confiança, enquanto o 4° quartil inclui as regras de menor suporte. Cada série, que corresponde a um país, recebe quatro valores percentuais correspondentes às quantidades percentuais de regras (do seu respectivo total de regras) nos quartis.

Utilizando as distribuições nos quartis, aplicou-se o algoritmo de agrupamento *K-Medoids*[8] com a função de distância euclidiana e o número de grupos variando de 2 a 100, e escolheu-se o número de grupos utilizando o valor da silhueta. O primeiro pico da silhueta após a queda inicial dos valores foi definido como o número adequado de grupos. Avaliou-se os grupos quanto à sua distribuição entre o 1° e o 2° quartis e entre o 1° e o 4° quartis, apresentando também as siglas dos países em cada cor de grupo na Figura 35 a Figura 37.

Nos gráficos, cada grupo apresenta uma cor e um símbolo, que não têm correspondência entre períodos, e o *medoid* do grupo é o país com sigla indicada no gráfico e de posição marcada por um ícone maior que os demais integrantes do grupo. O nome do país pode ser encontrado através da sua sigla ISO no Apêndice A.

A Figura 35 apresenta as distribuições dos países para o primeiro período, que estende-se de 1996 a 2004 e cobre a Crise da Bolha de Internet. Nota-se uma relação diretamente proporcional entre a quantidade de ocorrências no primeiro e no segundo quartis, Figura 35a, indicando que o país que possui mais ocorrências percentualmente no primeiro quartil, também terá mais ocorrências no segundo quartil. Já entre o primeiro e o último quartil, Figura 35b, a relação é inversamente proporcional, de forma que quanto maior o percentual de ocorrências do país em regras do primeiro quartil, menor o percentual de ocorrências no último.

O grupo de países que possui a maior quantidade percentual de ocorrências no primeiro e segundo quartil e menor quantidade no último quartil possui como *medoid* a Jordânia e contém 24 países integrantes. Os países integrantes desse grupo (Figura 35c) são os que melhor seguem a tendência mundial do período, já que possuem maior percentual de regras no 1° quartil, e estão entre eles: Estados Unidos, China e Emirados Árabes, países de atual economia forte. O grupo com maior quantidade percentual de regras no último quartil, também com 24 países integrantes, engloba Bolívia, Argentina e Brasil. A Figura 35 apresenta os grupos gerados no período de 1996 a 2004 respeitando as cores da Figura 35a e Figura 35b.

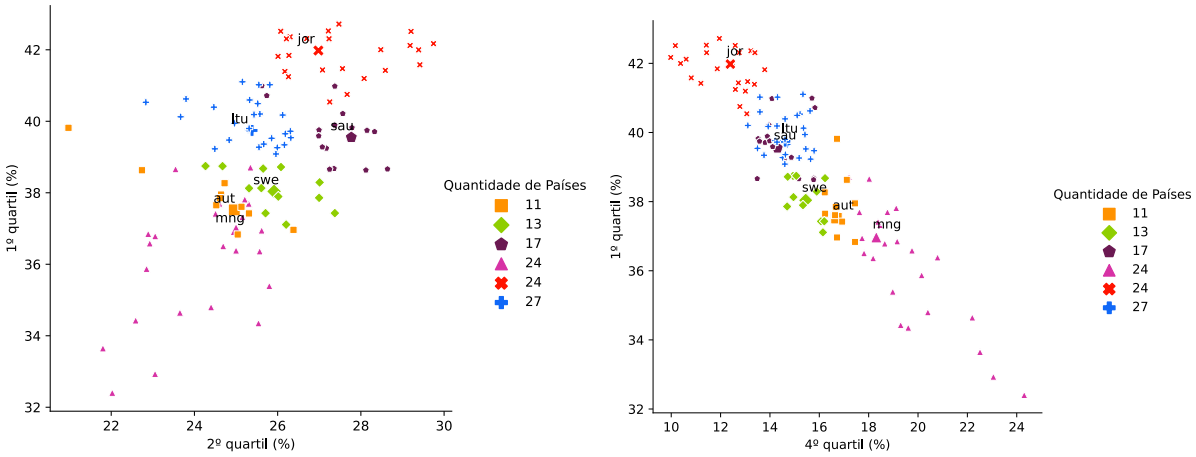
As principais regras do período envolvem aumento da importação, exportação e PIB. A regra com maior suporte é $([EXP,I] \Rightarrow [IMP,I], \Delta t = 0)$, $sup = 15,54$ $conf = 86,11$, que pode ser interpretada como: durante o período de 1996 a 2004, verifica-se um aumento na importação quando ocorre um aumento na exportação no mesmo ano com a confiança de 86%. Trata-se do comportamento mais forte do período e indica que a importação aumenta junto com a exportação, o que tipicamente é um indicador de crescimento econômico. Outra regra forte verificada é o aumento do PIB com o aumento da importação, que também reflete crescimento econômico no período.

[8] *K-Medoids* <https://scikit-learn-extra.readthedocs.io/en/stable/generated/sklearn_extra.cluster.KMedoids.html>

Figura 35 – Avaliação dos países nos quartis de regras de 1996 a 2004.

(a) 1° Quartil vs. 2° Quartil.

(b) 1° Quartil vs. 4° Quartil.



(c) Grupos de Países.

Fonte: Elaborada pela autora.

A partir da análise das regras genéricas, encontrou-se as regras menos comuns no período com a aplicação do corte. São regras que abrangem o decréscimo de todas as variáveis (importação, exportação, ECI e PIB). Logo, a queda em qualquer variável não é um comportamento comum do período. Um exemplo de regra com baixo suporte e confiança alta é $([PIB,D] \Rightarrow [IMP,I][EXP,I], \Delta t = 5)$, $sup = 3,11, conf = 88,15$. Essa regra indica que o aumento na importação e exportação 5 anos após o decréscimo no PIB é um comportamento pouco frequente (suporte baixo), mas geralmente é o que ocorre 5 anos após uma queda no PIB.

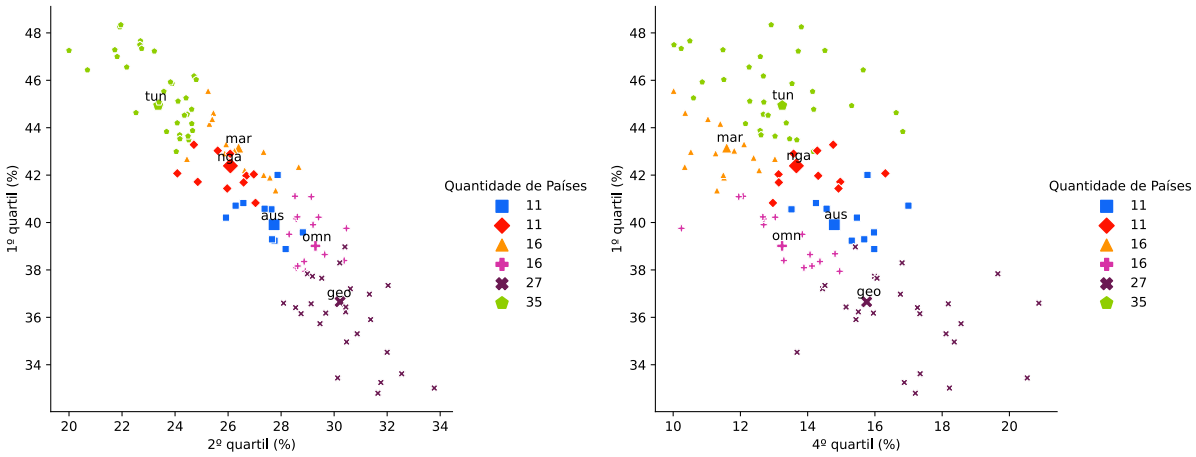
Embora tenha ocorrido a Crise da Bolha da Internet, o comportamento predominante do período é de crescimento econômico. A mineração da regra $([PIB,D] \Rightarrow [IMP,I][EXP,I], \Delta t = 5)$ pelo conjunto de dados desse período, contudo, permite verificar a retomada econômica,

ocorrendo em países como Japão e Canadá em 1998, Brasil em 1999, Rússia em 1998 e 1999, período da crise.

Figura 36 – Avaliação dos países nos quartis de regras de 2005 a 2012.

(a) 1° Quartil vs. 2° Quartil.

(b) 1° Quartil vs. 4° Quartil.



(c) Grupos de Países.

Fonte: Elaborada pela autora.

No período que cobre a Grande Recessão Econômica (2005-2012) observa-se uma relação inversa entre o percentual de ocorrência dos países no primeiro e segundo quartis, conforme Figura 36a. Já a relação apresentada entre o primeiro e último quartis na Figura 36b mantém-se inversamente proporcional como verificado no primeiro período, porém com um comportamento mais difuso.

O maior grupo é o que possui países com valores percentuais mais altos no primeiro quartil e mais baixos no segundo. Entre eles estão Peru, Equador e Portugal, visualizados na Figura 36c que contém as siglas de todos os países de cada grupo no período de 2005 a 2012. Ao analisar a distribuição entre o 1° e o 4° quartil, nota-se uma alta dispersão nesse grupo, sendo o

grupo com o *medoid* Marrocos que apresenta melhor a relação inversamente proporcional entre o primeiro e o último quartil. O Brasil encontra-se junto com Argentina, Chile e Venezuela, no grupo com menor percentual de ocorrências no primeiro quartil e maior percentual no 4° quartil.

As regras principais do período novamente envolvem aumento na importação, exportação e PIB, indicando a tendência mundial crescente para essas variáveis. A regra mais forte do período é $([IMP,I] \Rightarrow [PIB,I], \Delta t = 0)$, $sup = 18,84, conf = 92,62$ que sugere que o aumento da importação é seguido por um aumento do PIB no mesmo ano com alta frequência. Outra regra muito forte possui os mesmos padrões nas mesmas variáveis, mas com o período de 1 ano entre a ocorrência do antecedente e do consequente.

Nas regras com menor suporte do período, verifica-se o relacionamento entre decréscimo da importação, exportação e PIB em conjunto. Por exemplo, a regra $([IMP,D][PIB,D] \Rightarrow [EXP,D], \Delta t = 0)$, $sup = 3,48, conf = 94,78$ indica que o decréscimo da importação e do PIB são acompanhados do decréscimo da exportação no mesmo ano com uma baixa frequência mas alta certeza. Esse comportamento ocorreu em vários países entre 2009 e 2012, por exemplo nos Estados Unidos em 2009, Alemanha em 2009 e 2012 e no Brasil também nesses dois anos. As ocorrências da regra coincidem com a Grande Recessão, reafirmando a correspondência das regras com o comportamento econômico esperado.

A Figura 37 refere-se ao período econômico mais difuso (2013-2019), com o maior número de regras originalmente geradas e a menor quantidade de regras distintas após a aplicação do corte $sup_{min} = 3, conf_{min} = 50$. A quantidade maior de grupos indicada pelo método utilizando a silhueta reforça essa característica do período.

A Figura 37a apresenta a maior dispersão no percentual do 1° quartil de todos os períodos avaliados. Além disso, no período pré-covid verifica-se os maiores percentuais de composição do 4° quartil, conforme a Figura 37b. Esse comportamento ocorre devido ao aumento da disparidade entre as economias, com países deixando de seguir a tendência mundial de crescimento e especialização econômica.

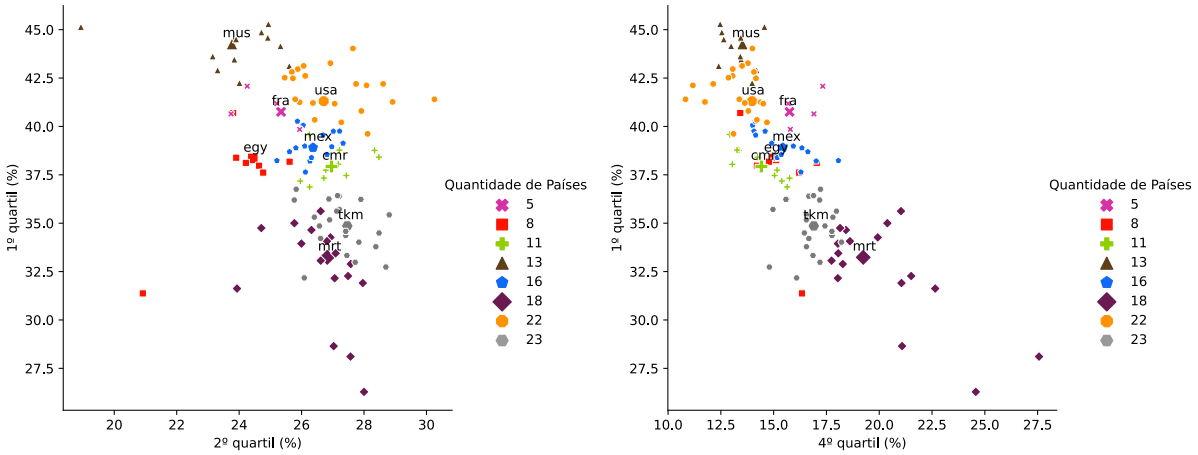
Entre as principais regras do período, verifica-se uma alta frequência de regras que abrangem a queda do ECI, indicando uma tendência mundial de diminuição da complexidade econômica dos países e que coincide com o comportamento de alta especialização na cadeia de produção e redução dos parceiros econômicos verificado no cenário econômico (UNCTAD, 2023). As regras de aumento na importação, exportação e PIB ainda possuem altos valores de suporte, mas tanto o suporte quanto a confiança dessas regras é menor em relação aos períodos anteriores.

As regras que representam os comportamentos mais incomuns do período incluem aumento no ECI frequentemente associado ao decréscimo das demais variáveis. No entanto, observa-se a regra de baixo suporte $([EXP,I][ECI,I] \Rightarrow [PIB,I], \Delta t = 2)$, $sup = 3,16, conf = 70$ que ocorre nos dados da China em 2014 e 2017, Japão em 2016 e Canadá em 2017, indicando

Figura 37 – Avaliação dos países nos quartis de regras de 2013 a 2019.

(a) 1° Quartil vs. 2° Quartil.

(b) 1° Quartil vs. 4° Quartil.



(c) Grupos de Países.

Fonte: Elaborada pela autora.

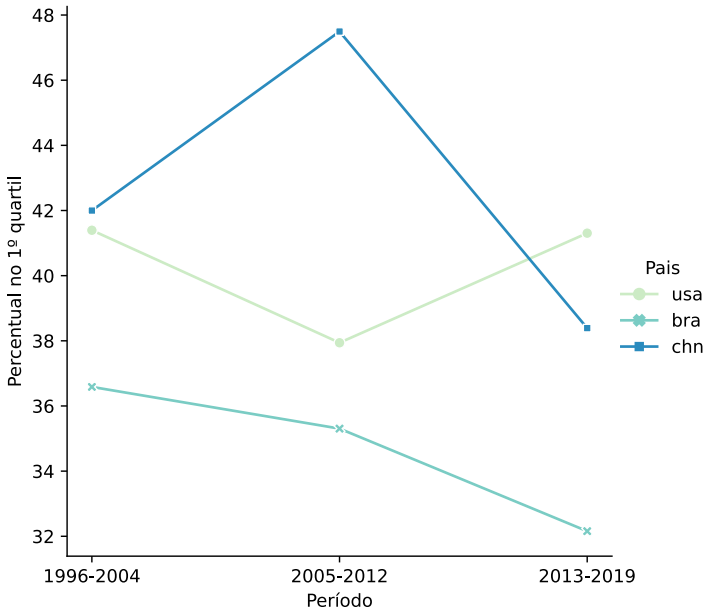
que 2 anos após o aumento na importação e no ECI, verificou-se o crescimento do PIB nesses países, que corresponde com o comportamento econômico esperado de aumento do PIB devido ao aumento da capacidade produtiva e do volume de exportação do país.

A análise quantitativa sobre períodos de interesse usando quartis permitiu verificar comportamentos econômicos esperados em cada período. No período que abrange a Crise da Bolha de Internet localiza-se regras que indicam a retomada econômica. Durante a Grande Recessão Econômica há regras confiáveis de queda nos índices econômicos e o período pré-covid apresenta padrões de aumento do ECI que indicam especialização das economias.

Para finalizar a análise semântica, avaliou-se o percentual de ocorrências no 1° quartil dos seguintes países: Estados Unidos, Brasil e China na Figura 38. Essa avaliação indica com qual

intensidade o país segue a tendência mundial, que é ditado pelas regras mais fortes pertencentes majoritariamente ao primeiro quartil.

Figura 38 – Evolução temporal do percentual de regras no 1° quartil dos Estados Unidos, Brasil e China.



Fonte: Elaborada pela autora.

Tanto os Estados Unidos quanto a China apresentam aproximadamente 42% das ocorrências do período pertencentes ao primeiro quartil de 1996 a 2004. Nesse período, a principal tendência mundial era de crescimento na importação, na exportação e no PIB, que correspondem com o comportamento apresentado por esses países durante o período. No Brasil, o percentual foi de aproximadamente 37%, demonstrando que o crescimento do país não foi tão acentuado.

De 2005 a 2012 a tendência de crescimento na importação, exportação e PIB é mantida nas regras e verifica-se um grande aumento do percentual no 1° quartil da China, indicando através das ocorrências nas regras, o seu forte crescimento. O Estados Unidos apresenta uma queda que indica um desacordo com a tendência mundial demonstrada pelas regras. Esse comportamento verificado pelas regras extensas coincide com o esperado devido à Grande Recessão Econômica de 2009. O Brasil, também amplamente afetado pela crise mundial, apresenta queda no percentual.

O último período possui comportamento mais difuso entre as regras e é, portanto, o período mais difícil de análise. Enquanto os Estados Unidos apresentam crescimento na correspondência com a tendência mundial, China e Brasil apresentaram queda. Embora não se observe uma tendência mundial bem definida, as regras ainda indicam um crescimento na importação, exportação e PIB, que verifica-se para os Estados Unidos. Na China, ocorre um aumento na complexidade econômica durante o período, que se mostra contrário ao comportamento mundial. Já o Brasil possui um comportamento mais estável no período, sem crescimento econômico destacado.

Respondendo a terceira questão de pesquisa, as regras encontradas na mineração de séries do comércio internacional confirmam a tendência mundial de crescimento econômico. Além disso, na análise dos períodos de interesse é possível verificar regras confiáveis de países como Estados Unidos, Brasil e China que são coerentes com o estado da economia no período.

4.6 Considerações Finais

O eTRUMiner foi aplicado sobre dados econômicos de fontes variadas que constituem séries temporais heterogêneas e incompletas. Os dados abrangem importação, exportação, ECI e PIB de 232 países e territórios com duração de 1996 a 2020, periodicidade anual e percentual relativo de observações em cada variável variando entre 97% e 99%. Do conjunto de dados original derivou-se o conjunto de dados homogêneo contendo 116 séries completas com duração de 24 anos.

Na Seção 4.2 avaliou-se os métodos de discretização implementados analisando a distribuição percentual das regras no suporte, na confiança, e o volume de regras geradas. Mostrou-se que discretização decis não é adequada para o conjunto de séries do comércio internacional, com elevado número de regras gerado e alto percentual de regras com suporte muito baixo. Para todos as discretizações avaliadas, verificou-se a alta presença de regras com baixo suporte e baixa confiança, indicando a necessidade da aplicação de corte nas métricas de avaliação. Verificou-se na Seção 4.3 que a delimitação de suporte mínimo elimina de forma mais efetiva as regras não relevantes, ou seja, mesmo com a aplicação de um corte baixo, grande parte das regras com suporte e confiança baixos são excluídas. Para as discretizações avaliadas, o corte no suporte indicado foi $sup_{min} \geq 2$, com o corte $conf_{min} \geq 50$ adicional nos métodos comportamental e SAX para seleção mais rigorosa.

O uso de conjuntos de dados com observações e variáveis faltantes mostra-se muitas vezes como única alternativa. Na Seção 4.4 verificou-se que o eTRUMiner é capaz de lidar com esse cenário sem a necessidade de tratamentos adicionais. O comportamento do volume da distribuição das regras no suporte e na confiança entre os conjuntos original e homogêneo apresentou alta similaridade, com suporte máximo apresentando variação máxima de 15% entre conjuntos. As regras retornadas por conjuntos com dados faltantes possui 100% de intersecção entre as regras e em alta coincidência nas métricas de avaliação. A aplicação de corte nas métricas de avaliação apresenta-se como extremamente eficaz para selecionar as regras fortes com alta semelhança.

Na análise semântica Seção 4.5, a avaliação de regras genéricas mostrou que a principal influência para obtenção de baixo suporte nas regras temporais é a característica temporal. A principal regra genérica do comércio internacional é $[IMP, I] \Rightarrow [PIB, I]$ que indica o aumento na importação seguido de um aumento no PIB. Na discretização quartis, o principal aumento dessa regra é de 25% na importação e no PIB, reforçando e detalhando a regra obtida da

discretização comportamental. Ao avaliar as regras temporais nos períodos econômicos de interesse, confirmou-se a coerência das regras obtidas com correspondência do perfil econômico mundial esperado.

O próximo capítulo apresenta as conclusões obtidas neste trabalho de mestrado, destacando os principais resultados atingidos. As questões introduzidas no Capítulo 1 são respondidas e discutidas, e possíveis trabalhos futuros são propostos com indicações de abordagens a serem tomadas.

CAPÍTULO

5

CONCLUSÃO

O presente trabalho de mestrado propôs uma solução e desenvolveu um algoritmo capaz de minerar regras temporais multivariadas, o eTRUMiner. Para tal, realizou-se um desenvolvimento teórico quanto à definição de transação, regra temporal, suporte e confiança. O eTRUMiner permite verificar o relacionamento entre diversas variáveis para diferentes intervalos temporais, podendo informar todas as ocorrências de cada regra nas séries. O algoritmo desenvolvido neste trabalho de mestrado, facilita a mineração de regras temporais de conjuntos de dados reais, pois permite a utilização de múltiplas fontes de dados e é capaz de lidar com observações e variáveis faltantes.

Algoritmos similares não tratam problemas existentes em conjuntos de dados reais como observações e variáveis faltantes e duração heterogênea entre as variáveis das séries analisadas. O eTRUMiner é capaz de minerar múltiplas séries temporais multivariadas heterogêneas e incompletas, retornando regras no formato curto e extenso. Além disso, as regras são compostas de duas ou mais variáveis.

Os métodos de discretização já implementados são o comportamental, decis, quartis e SAX. Os três primeiros métodos avaliam a variação entre observações consecutivas, sendo a discretização comportamental a que gera menos elementos discretizados distintos, enquanto a decis gera o maior volume mas os símbolos possuem informações mais detalhadas a respeito da variação entre observações. A implementação de novos métodos de discretização é facilitada permitindo a ampliação de sua área de aplicação.

O eTRUMiner gera regras multivariadas contendo o máximo de variáveis distintas possível. Na etapa de geração de regras, une-se os padrões de antecedentes frequentes aos padrões de consequentes frequentes desde que a intersecção de ocorrência desses padrões seja acima do mínimo pré-determinado. Esse processo é realizado para cada intervalo temporal individualmente e as regras geradas são ordenadas internamente para evitar geração de regras idênticas em duplicidade.

A avaliação das regras já implementada constitui-se do suporte e confiança, mas o eTRUMiner também aceita a adição de outros métodos de avaliação. Na avaliação do suporte de regras temporais, deve-se levar em consideração a subdivisão das transações em diversos intervalos temporais até o limite da janela temporal pré-determinado, o que reduz o valor máximo possível do suporte. O cálculo das métricas de avaliação é facilitado pelo armazenamento das ocorrências nas transações realizado pelo algoritmo, e as regras são retornadas acima dos cortes pré-determinados, ordenadas por suporte e confiança.

O eTRUMiner foi aplicado e avaliado sobre séries de importação, exportação, ECI e PIB de países e territórios com duração de 1996 a 2020. As discretizações comportamental, decis, quartis e SAX foram aplicadas ao conjunto de dados original e homogêneo. Analisou-se os valores adequados de cortes nas métricas de avaliação, o impacto de dados faltantes e principais regras do conjunto do comércio internacional.

Entre as discretizações avaliadas, a comportamental apresentou os maiores valores de suporte máximo, melhor distribuição do suporte e da confiança no percentual de regras distintas, gerando a menor dispersão de regras distintas. A discretização SAX, embora apresente bons resultados quanto ao volume de regras geradas, quanto à aplicação de medidas de corte e quanto aos valores das métricas de avaliação, não se mostra como a mais indicada, dado que compreensão dos padrões das regras são específicos às suas ocorrências, demandando uma interpretação mais individualizada. Os métodos decis e quartis são inadequados para esse conjunto de dados, sendo a discretização quartis indicada para uma análise auxiliar.

A aplicação do eTRUMiner sobre o conjunto de dados do comércio internacional indica que embora um conjunto de dados homogêneo e completo retorne um conjunto mais conciso de regras, com suporte maiores, é possível minerar regras temporais a partir de séries econômicas heterogêneas e incompletas. O impacto é aceitável, apresentando baixa penalização sobre as regras retornadas. Além disso, o uso de cortes nas métricas de avaliação surge como uma alternativa para selecionar as regras de interesse.

As regras retornadas da mineração de séries do comércio internacional apresentam coerência com o comportamento econômico mundial esperado de cada período avaliado, com a discretização quartis detalhando os aumentos em 25%, principalmente. Uma análise mais direcionada permite verificar comportamentos confiáveis específicos de países como Estados Unidos, Brasil e China através das regras extensas.

5.1 Contribuições Científicas

As contribuições científicas do presente trabalho de mestrado englobam concepção teórica da solução proposta para o problema tratado, desenvolvimento e implementação de um método capaz de minerar regras temporais com 2 ou mais variáveis de séries temporais heterogêneas incompletas, além da aplicação e avaliação em dados reais, e trabalhos publicados.

A definição de transação para regra temporal multivariada e a própria regra temporal multivariada com mais de 2 variáveis foram novas propostas presentes neste trabalho. Cada transação já engloba a informação temporal que será a característica temporal da regra e possui todos os padrões que podem compor o antecedente e o consequente da regra temporal. Com a redefinição de regra temporal multivariada genérica, propõe-se uma avaliação de suporte e confiança adaptados, próprios para essas regras temporais.

O algoritmo eTRUMiner (*extended Temporal RUles Miner*), desenvolvido durante este trabalho de mestrado, produz regras temporais contendo duas ou mais variáveis distintas a partir de séries temporais multivariadas heterogêneas incompletas. A implementação em C++ inclui diversos métodos de discretização e permite facilmente integrar novos métodos de discretização, além de outras métricas de avaliação. O algoritmo desenvolvido pode informar todas as ocorrências nas séries de cada regra retornada e possui estratégias para redução de complexidade temporal e espacial como uso da janela temporal, ordem pré-definida entre variáveis, tempo inicial constante entre padrões do antecedente/consequente, armazenamento de ocorrências dos padrões, suporte e confiança mínimos.

A mineração de regras temporais do conjunto de dados do comércio internacional retornou regras condizentes com o perfil econômico esperado, frequentemente verificando-se crescimento nas variáveis de importação, exportação e PIB. O número de regras temporais geradas é elevado, mas o corte no suporte e na confiança mostram-se eficazes na redução do volume e seleção de regras fortes. Verifica-se para esse conjunto de dados, que a discretização comportamental é a mais adequada e permite fácil compreensão das regras. Regras derivadas de comportamentos mais específicos como crises e queda do Índice de Complexidade Econômica são identificadas e podem auxiliar na análise desses eventos.

Os trabalhos publicados com os resultados obtidos durante a pesquisa de mestrado estão listados abaixo.

- KARASAWA, E.; SOUSA, E. Truminer: Mineração de regras temporais em bases de séries multivariadas e heterogêneas. In: **Anais do XXXVII Simpósio Brasileiro de Bancos de Dados.** Porto Alegre, RS, Brasil: SBC, 2022. p. 403–408. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/21827>>.

- KARASAWA, E. G.; SOUSA, E. P. M. Mining temporal rules from heterogeneous multivariate time series. **Journal of Information and Data Management**, v. 14, n. 2, Dec. 2023. Disponível em: <https://sol.sbc.org.br/journals/index.php/jidm/article/view/3232>.

O *short paper "Truminer: Mineração de regras temporais em bases de séries multivariadas e heterogêneas"* foi apresentado e publicado no 37º Simpósio Brasileiro de Banco de Dados (SBBD 2022), recebendo o prêmio de *menção honrosa*. O trabalho *"Mining temporal rules from*

heterogeneous multivariate time series" trata-se de uma extensão dos resultados publicados no SBBD 2022.

5.2 Trabalhos Futuros

A pesquisa realizada durante este trabalho de mestrado permitiu iniciar uma nova abordagem para a mineração de dados sobre o comércio internacional. Durante o estudo, verificou-se possíveis trabalhos futuros brevemente propostos a seguir:

- Mineração com múltiplas discretizações – Dado que o processo de discretização é aplicado sobre cada série univariada, é possível aplicar o método de discretização mais apropriado para cada variável analisada, obtendo regras mais delimitadas a critério da análise específica desejada.
- Novas métricas de avaliação – A aplicação de novas métricas de avaliação permite selecionar de forma mais concisa as regras desejadas, podendo ser agregadas às já existentes para avaliações mais diretas.
- Mineração de outros conjuntos de dados – Avaliação do eTRUMiner quando aplicado a conjuntos de dados em outras áreas de conhecimento, por exemplo, dados agrometeorológicos. Espera-se que a versatilidade do algoritmo possibilite a descoberta de conhecimento a partir de outros domínios de aplicação.
- Avaliação aprofundada com auxílio de especialista – A análise auxiliada pelo especialista pode permitir a descoberta de novos conhecimentos a respeito do comércio internacional, bem como direcionar a própria mineração de regras.

REFERÊNCIAS

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. Association for Computing Machinery, 1993. Disponível em: <https://doi.org/10.1145/170035.170072>. Citado nas páginas 29, 36 e 37.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. Morgan Kaufmann Publishers Inc., 1994. Citado na página 37.

BERZAL, F.; BLANCO, I.; SÁNCHEZ, D.; VILA, M.-A. Measuring the accuracy and interest of association rules: A new framework. IOS Press, 2002. Citado na página 36.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. **Time series analysis: forecasting and control**. [S.l.]: John Wiley & Sons, 2015. Citado na página 35.

CHEN, X.; PETROUNIAS, I. Discovering temporal association rules: Algorithms, language and system. 2000. Disponível em: <https://doi.org/10.1109/ICDE.2000.839423>. Citado na página 37.

DAR, Q.; DAR, G. F.; MA, J.-H.; AHN, Y.-H. Visualization, economic complexity index, and forecasting of south korea international trade profile: a time series approach. **Journal of Korea Trade**, 2020. Citado na página 31.

DAS, G.; LIN, K.-I.; MANNILA, H.; RENGANATHAN, G.; SMYTH, P. Rule discovery from time series. AAAI Press, 1998. Citado na página 41.

EUROSTAT. **Statistics on the Trading of Goods-User Guide**. Office for the Official Publications of the European Communities Luxembourg, 2006. Disponível em: <https://ec.europa.eu/eurostat/ramon/statmanuals/files/KS-BM-05-001-EN.pdf>. Acesso em: 03/03/2021. Citado na página 30.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, 1996. Disponível em: <https://doi.org/10.1609/aimag.v17i3.1230>. Citado na página 29.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery and data mining**. USA: American Association for Artificial Intelligence, 1996. Citado na página 29.

FOURNIER-VIGER, P.; FAGHIHI, U.; NKAMBOU, R.; NGUIFO, E. M. Cmrules: Mining sequential rules common to several sequences. **Knowledge-Based Systems**, 2012. Disponível em: <https://doi.org/10.1016/j.knosys.2011.07.005>. Citado na página 42.

GOMEZ-GONZALEZ, J. E.; URIBE, J. M.; VALENCIA, O. M. Does economic complexity reduce the probability of a fiscal crisis? **World Development**, 2023. Disponível em: <https://doi.org/10.1016/j.worlddev.2023.106250>. Citado nas páginas 30 e 31.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011. Citado nas páginas 29 e 37.

HARMS, S. K.; DEOGUN, J. S. Sequential association rule mining with time lags. **Journal of Intelligent Information Systems**, 2004. Disponível em: <https://doi.org/10.1023/A:1025824629047>. Citado nas páginas 31, 38, 42 e 44.

KARASAWA, E.; SOUSA, E. Truminer: Mineração de regras temporais em bases de séries multivariadas e heterogêneas. In: . Porto Alegre, RS, Brasil: SBC, 2022. p. 403–408. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/21827>. Citado na página 33.

KARASAWA, E. G.; SOUSA, E. P. M. Mining temporal rules from heterogeneous multivariate time series. **Journal of Information and Data Management**, v. 14, n. 2, Dec. 2023. Disponível em: <https://sol.sbc.org.br/journals/index.php/jidm/article/view/3232>. Citado na página 33.

KEOGH, E.; CHAKRABARTI, K.; PAZZANI, M.; MEHROTRA, S. Dimensionality reduction for fast similarity search in large time series databases. **Knowledge and Information Systems**, 2001. Disponível em: <https://doi.org/10.1007/PL00011669>. Citado na página 40.

KEOGH, E.; LIN, J. Clustering of time-series subsequences is meaningless: implications for previous and future research. **Knowledge and information systems**, 2005. Disponível em: <https://doi.org/10.1007/s10115-004-0172-7>. Citado na página 42.

KEOGH, E. J.; PAZZANI, M. J. Scaling up dynamic time warping for datamining applications. In: . Association for Computing Machinery, 2000. Disponível em: <https://doi.org/10.1145/347090.347153>. Citado na página 40.

LIN, J.; KEOGH, E.; LONARDI, S.; CHIU, B. A symbolic representation of time series, with implications for streaming algorithms. In: . Association for Computing Machinery, 2003. Disponível em: <https://doi.org/10.1145/882082.882086>. Citado nas páginas 40 e 41.

LIN, J.; KEOGH, E.; WEI, L.; LONARDI, S. Experiencing sax: a novel symbolic representation of time series. **Data Mining and Knowledge Discovery**, 2007. Disponível em: <https://doi.org/10.1007/s10618-007-0064-z>. Citado na página 40.

MITSA, T. **Temporal data mining**. [S.l.]: CRC Press, 2010. Citado nas páginas 35 e 39.

MORETTIN, P. A.; TOLOI, C. **Análise de séries temporais**. [S.l.]: Blucher, 2006. Citado na página 35.

NAM, H.; LEE, K.; LEE, D. Identification of temporal association rules from time-series microarray data set: temporal association rules. In: . Association for Computing Machinery, 2008. Disponível em: <https://doi.org/10.1145/1458449.1458457>. Citado nas páginas 42, 43 e 45.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. d. **Mineração de dados**. [S.l.: s.n.], 2003. Citado na página 29.

ROMANI, L. A. S.; AVILA, A. M. H. de; CHINO, D. Y. T.; ZULLO, J.; CHBEIR, R.; TRAINA, C.; TRAINA, A. J. M. A new time series mining approach applied to multitemporal remote sensing imagery. 2013. Disponível em: <https://doi.org/10.1109/TGRS.2012.2199501>. Citado na página 44.

ROMANI, L. A. S.; AVILA, A. M. H. de; ZULLO, J.; CHBEIR, R.; TRAINA, C.; TRAINA, A. J. M. Clearminer: a new algorithm for mining association patterns on heterogeneous time series from climate data. Association for Computing Machinery, 2010. Disponível em: <https://doi.org/10.1145/1774088.1774275>. Citado nas páginas 31, 38, 42, 43, 45, 48 e 49.

SEGURA-DELGADO, A.; GACTO, M. J.; ALCALÁ, R.; ALCALÁ-FDEZ, J. Temporal association rule mining: An overview considering the time variable as an integral or implied component. Wiley Online Library, 2020. Disponível em: <https://doi.org/10.1002/widm.1367>. Citado nas páginas 29 e 38.

SMYTH, P.; GOODMAN, R. M. An information theoretic approach to rule induction from databases. IEEE, 1992. Disponível em: <https://doi.org/10.1109/69.149926>. Citado na página 36.

TACCHELLA, A.; MAZZILLI, D.; PIETRONERO, L. A dynamical systems approach to gross domestic product forecasting. Nature Publishing Group UK London, 2018. Disponível em: <https://doi.org/10.1038/s41567-018-0204-y>. Citado nas páginas 30 e 31.

TAN, T.-F.; WANG, Q.-G.; LI, X.; HUANG, J.; PHANG, T.-H. Temporal association rule mining: With application to us stock market. **Transactions on Machine Learning and Artificial Intelligence**, 2015. Disponível em: <https://doi.org/10.14738/tmlai.35.1051>. Citado na página 41.

UNCTAD. **International Trade Statistics Yearbook 2019**. United Nations Publications, 2019. Disponível em: <https://digitallibrary.un.org/record/3900353?ln=es>. Acesso em: 03/03/2021. Citado na página 29.

____. **Trade and Development Report 2020**. United Nations Publications, 2020. Disponível em: <https://unctad.org/system/files/official-document/tdr2020_en.pdf>. Acesso em: 15/12/2023. Citado na página 70.

____. **Trade and Development Report 2023**. United Nations Publications, 2023. Disponível em: <https://unctad.org/system/files/official-document/tdr2023_en.pdf>. Acesso em: 16/10/2023. Citado nas páginas 30, 71, 97 e 103.

WANG, L.; GUI, L.; XU, P. Incremental sequential patterns for multivariate temporal association rules mining. Pergamon Press, Inc., 2022. Disponível em: <https://doi.org/10.1016/j.eswa.2022.118020>. Citado na página 42.

WANG, L.; MENG, J.; XU, P.; PENG, K. Mining temporal association rules with frequent itemsets tree. Elsevier, 2018. Disponível em: <https://doi.org/10.1016/j.asoc.2017.09.013>. Citado na página 42.

WORLDBANK; OECD. **GDP**. World Bank Group, 2023. Disponível em: <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2022&start=1996&view=chart>. Acesso em: 15/12/2023. Citado na página 72.

ZAKI, M. J. Spade: An efficient algorithm for mining frequent sequences. Springer, 2001. Disponível em: <https://doi.org/10.1023/A:1007652502315>. Citado na página 57.

ZARNOWITZ, V.; OZYILDIRIM, A. Time series decomposition and measurement of business cycles, trends and growth cycles. **Journal of Monetary Economics**, 2006. Disponível em: <https://doi.org/10.1016/j.jmoneco.2005.03.015>. Citado na página 67.

APÊNDICE

A

TERRITÓRIOS E SIGLA ISO

Tabela 7 – Territórios Presentes nos Datasets e suas Respectivas Siglas (A-B).

Território	**Sigla**
Afeganistão	AFG
Albânia	**ALB**
Alemanha	**DEU**
Andorra	AND
Angola	**AGO**
Anguilla	AIA
Antilhas Holandesas	ANT
Antígua and Barbuda	ATG
Argentina	**ARG**
Argélia	**DZA**
Armênia	**ARM**
Aruba	ABW
Arábia Saudita	**SAU**
Austrália	**AUS**
Azerbaijão	**AZE**
Bahamas	BHS
Bahrein	**BHR**
Bangladesh	**BGD**
Barbados	BRB
Belize	BLZ
Benin	BEN
Bermudas	BMU
Bielorrússia	**BLR**
Bolívia	**BOL**
Botsuana	BWA
Brasil	**BRA**
Brunei	BRN
Bulgária	**BGR**

Tabela 8 – Territórios Presentes nos Datasets e suas Respectivas Siglas (C-E).

Território	**Sigla**
Burquina Faso	**BFA**
Burundi	BDI
Butão	BTN
Bélgica	**BEL**
Bósnia e Herzegovina	**BIH**
Cabo Verde	CPV
Camarões	**CMR**
Cambodja	**KHM**
Canadá	**CAN**
Catar	**QAT**
Cazaquistão	**KAZ**
Chade	TCD
Chile	**CHL**
China	**CHN**
Chipre	**CYP**
Cisjordânia	WBG
Colômbia	**COL**
Comores	COM
Congo	COG
Coreia do Norte	PRK
Coreia do Sul	KOR
Costa Rica	**CRI**
Costa do Marfim	CIV
Croácia	**HRV**
Cuba	CUB
Curaçao	CUW
Dinamarca	**DNK**
Djibuti	DJI
Dominica	DMA
Egito	**EGY**
El Salvador	**SLV**
Emirados Árabes Unidos	**ARE**
Equador	**ECU**
Eritreia	ERI
Eslováquia	**SVK**
Eslovênia	**SVN**
Espanha	**ESP**
Essuatíni	SWZ
Estados Federados da Micronésia	FSM
Estados Unidos	**USA**
Estônia	**EST**
Etiópia	**ETH**

Tabela 9 – Territórios Presentes nos Datasets e suas Respectivas Siglas (F-J).

Território	**Sigla**
Fiji	FJI
Filipinas	**PHL**
Finlândia	**FIN**
França	**FRA**
Gabão	**GAB**
Gana	GHA
Geórgia	**GEO**
Gibraltar	GIB
Granada	GRD
Groenlândia	GRL
Grécia	**GRC**
Guam	GUM
Guatemala	**GTM**
Guiana	GUY
Guiné	GIN
Guiné Equatorial	GNQ
Guiné-Bissau	GNB
Gâmbia	GMB
Haiti	HTI
Honduras	**HND**
Hong Kong	HKG
Hungria	**HUN**
Ilha Christmas	CXR
Ilha Norfolk	NFK
Ilhas Cayman	CYM
Ilhas Cocos	CCK
Ilhas Cook	COK
Ilhas Malvinas	FLK
Ilhas Marianas do Norte	MNP
Ilhas Marshall	MHL
Ilhas Maurício	**MUS**
Ilhas Pitcairn	PCN
Ilhas Salomão	SLB
Ilhas Virgens Britânicas	VGB
Indonésia	**IDN**
Iraque	IRQ
Irlanda	**IRL**
Irã	**IRN**
Islândia	ISL
Israel	**ISR**
Itália	**ITA**
Iêmen	**YEM**
Jamaica	**JAM**
Japão	**JPN**
Jordânia	**JOR**

Tabela 10 – Territórios Presentes nos Datasets e suas Respectivas Siglas (K-O).

Território	**Sigla**
Kiribati	KIR
Kosovo	UVK
Kuwait	**KWT**
Laos	**LAO**
Lesoto	LSO
Letônia	**LVA**
Libéria	LBR
Lituânia	**LTU**
Luxemburgo	LUX
Líbano	LBN
Líbia	**LBY**
Macau	MAC
Macedônia do Norte	**MKD**
Madagascar	MDG
Malawi	**MWI**
Maldivas	MDV
Mali	MLI
Malta	MLT
Malásia	**MYS**
Marrocos	**MAR**
Mauritânia	**MRT**
Mayotte	MYT
Mianmar	MMR
Moldávia	**MDA**
Mongólia	**MNG**
Montenegro	MNE
Montserrat	MSR
Moçambique	**MOZ**
México	**MEX**
Namíbia	NAM
Nauru	NRU
Nepal	NPL
Nicarágua	**NIC**
Nigéria	**NGA**
Niue	NIU
Noruega	**NOR**
Nova Caledônia	NCL
Nova Zelândia	**NZL**
Níger	NER
Omã	**OMN**

Tabela 11 – Territórios Presentes nos Datasets e suas Respectivas Siglas (P-S).

Território	**Sigla**
Palau	PLW
Palestina	PSE
Panamá	**PAN**
Papua-Nova Guiné	**PNG**
Paquistão	**PAK**
Paraguai	**PRY**
Países Baixos Caribenhos	BES
Países-Baixos	**NLD**
Peru	**PER**
Polinésia Francesa	PYF
Polônia	**POL**
Porto Rico	PRI
Portugal	**PRT**
Quirguistão	**KGZ**
Quênia	**KEN**
Reino Unido	**GBR**
República Centro-Africana	CAF
República Democrática do Congo	**COD**
República Dominicana	**DOM**
Romênia	**ROU**
Ruanda	RWA
Rússia	**RUS**
Saint-Pierre e Miquelon	SPM
Samoa	WSM
Samoa Americana	ASM
San Marino	SMR
Santa Helena, Ascensão e Tristão da Cunha	SHN
Santa Lúcia	LCA
Senegal	SEN
Serra Leoa	SLE
Seychelles	SYC
Singapura	**SGP**
Somália	SOM
Sri Lanka	**LKA**
Sudão	SDN
Sudão do Sul	SSD
Suriname	SUR
Suécia	**SWE**
Suíça	**CHE**
São Bartolomeu	BLM
São Cristóvão e Névis	KNA
São Martinho	SXM

Tabela 12 – Territórios Presentes nos Datasets e suas Respectivas Siglas (S-Í).

Território	**Sigla**
São Tomé e Príncipe	STP
São Vicente e Granadinas	VCT
Sérvia	SRB
Sérvia e Montenegro	SCG
Síria	SYR
Tailândia	**THA**
Taiwan	TWN
Tajiquistão	**TJK**
Tanzânia	**TZA**
Tchéquia	**CZE**
Terras Austrais e Antárticas Francesas	ATF
Território Britânico do Oceano Índico	IOT
Timor-Leste	TLS
Togo	**TGO**
Tonga	TON
Toquelau	TKL
Trindade e Tobago	**TTO**
Tunísia	**TUN**
Turcomenistão	**TKM**
Turks e Caicos	TCA
Turquia	**TUR**
Tuvalu	TUV
Ucrânia	**UKR**
Uganda	**UGA**
Uruguai	**URY**
Uzbequistão	**UZB**
Vanuatu	VUT
Venezuela	**VEN**
Vietnã	**VNM**
Wallis e Futuna	WLF
Zimbábue	ZWE
Zâmbia	**ZMB**
África do Sul	**ZAF**
Áustria	**AUT**
Índia	**IND**


ICMC
USP
SÃO CARLOS