UNIVERSIDADE CANDIDO MENDES – UCAM
PROGRAMA DE PÓS-GRADUAÇÃO EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL
CURSO DE MESTRADO EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL

Pedro Sant’ana Bastos da Silva

MINERAÇÃO DE DADOS NO COMBATE AOS CARTÉIS

CAMPOS DOS GOYTACAZES
Fevereiro de 2020

UNIVERSIDADE CANDIDO MENDES – UCAM
PROGRAMA DE PÓS-GRADUAÇÃO EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL
CURSO DE MESTRADO EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL

Pedro Sant'ana Bastos da Silva

MINERAÇÃO DE DADOS NO COMBATE AOS CARTÉIS

Dissertação apresentada ao Programa de Pós-Graduação em Pesquisa Operacional e Inteligência Computacional, da Universidade Candido Mendes – Campos/RJ, para obtenção do grau de MESTRE EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL.

Orientador: Prof. Dr. Ítalo de Oliveira Matias.

CAMPOS DOS GOYTACAZES
Fevereiro de 2020

Catalogação na Fonte

Preparada pela Biblioteca da **UCAM – CAMPOS** **029/2020**

Silva, Pedro Sant'ana Bastos da.

Mineração de dados no combate aos cartéis. / Pedro Sant'ana Bastos da Silva – 2020.
80 f.: il.

Orientador: Ítalo de Oliveira Matias.

Dissertação de Mestrado em Pesquisa Operacional e Inteligência Computacional – Universidade Candido Mendes – Campos. Campos dos Goytacazes, RJ, 2020.
Referências: f. 78-80.

1. Licitações públicas. 2. Mineração de dados. I. Matias, Ítalo de Oliveira, orient. II. Universidade Candido Mendes – Campos. III. Título.

CDU – 351.712:004.62

Bibliotecária Responsável: Flávia Mastrogirolamo CRB 7ª-6723

PEDRO SANT'ANA BASTOS DA SILVA

MINERAÇÃO DE DADOS NO COMBATE AOS CARTÉIS

Dissertação apresentada ao Programa de Pós-Graduação em Pesquisa Operacional e Inteligência Computacional, da Universidade Candido Mendes – Campos/RJ, para obtenção do grau de MESTRE EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL.

Aprovado em 14 de fevereiro de 2020.

BANCA EXAMINADORA

_____

Prof. Ítalo de Oliveira Matias, D.Sc. – Orientador
Universidade Cândido Mendes (UCAM)

_____

Prof. Aldo Shimoya, D.Sc.
Universidade Cândido Mendes (UCAM)

_____

Prof. Felipe Gonçalves Figueira, D.Sc.
Instituto Federal Fluminense (IFF)

CAMPOS DOS GOYTACAZES/RJ

2020

Dedico este trabalho aos meus pais, irmãos, familiares e amigos, por estarem sempre ao meu lado nos momentos difíceis e de alegria, e por me motivarem a evoluir e melhorar a cada dia como profissional e como pessoa.

AGRADECIMENTOS

Em primeiro lugar, quero agradecer a Deus, pois até aqui me ajudou o senhor, e sem sua vontade, este trabalho não estaria concluído.

Ao meu orientador Dr. Ítalo de Oliveira Matias, pela dedicação a mim concedida, paciência, e atenção nos ensinamentos pertinentes a este estudo.

À Universidade Candido Mendes (UCAM) e ao seu corpo docente, sempre muito atencioso e solícito para com as minhas demandas de conhecimento, me oferecendo uma oportunidade de aprendizado de excelência.

Aos meus pais, que estão sempre ao meu lado me dando todo o suporte necessário para prosseguir na minha caminhada profissional e pessoal, e que moldaram em cada detalhe a pessoa que sou hoje.

Aos meus irmãos e aos demais familiares, que sempre entenderam que minhas ausências para com eles eram por uma boa causa, a fim de concluir este mestrado e gozar futuramente de seus benefícios.

Aos meus companheiros de classe, que dividiram comigo horas e horas de aulas durante os fins de semana, sempre em um ambiente de cumplicidade e incentivo uns com os outros, gerando assim força para que cada amigo de turma pudesse completar sua caminhada.

E a todos que de alguma forma contribuíram para a conclusão deste mestrado.

RESUMO

MINERAÇÃO DE DADOS NO COMBATE AOS CARTÉIS

As licitações públicas são processos de compra que visam dar tratamento isonômico aos fornecedores, de modo a criar um ambiente justo de competição entre eles, que estimule a queda dos preços pagos pelo governo. Entretanto, existe uma fraude denominada cartel, que frustra com os objetivos governamentais, ocasiona o aumento dos valores dos contratos, e beneficia aos particulares que agem em conluio. Trata-se de um acordo prévio de preços entre os participantes da licitação, que geralmente atuam em conjunto nesses processos de compra, e se revezam na posição vencedora. Em vista disso, esta dissertação teve como objetivo propor uma metodologia de combate a esta fraude, visando subsidiar o trabalho dos auditores fiscais, fornecendo maior eficiência ao processo de fiscalização. Primeiramente foi mapeado o “estado da arte”, identificando quais são as estratégias usadas pela comunidade científica até o momento, com a identificação dos trabalhos e autores mais relevantes nesta linha de pesquisa. Uma vez definida a metodologia a ser aplicada, proposta em Silva (2011) e alterada por este trabalho, foi realizado um referencial teórico envolvendo todos os conceitos de Descoberta de Conhecimento em Base de Dados (DCBD) necessários ao entendimento deste estudo. Feito isso, o método foi apresentado e testado com um estudo de caso aplicado a uma base de dados real de licitações, disponível no sítio eletrônico da Controladoria Geral da União (CGU), e os resultados se mostraram promissores.

Palavras-chave: Mineração de Dados. DCBD. Licitações. Cartéis.

ABSTRACT

DATA MINING APPLIED IN FIGHTING CARTELS

The public bids are purchases processes that aim to provide isonomic treatment between the suppliers in order to create a fair competitive environment that encourages the falling of the prices paid by the government. However, there is a fraud called cartel, which frustrates the government objectives, increases the contract's values, and benefits the individuals who practice this collusion. This is a prior price agreement between the bidding participants, who generally act together in these purchase processes and take turns in the winning position. Therefore, this dissertation has the objective to propose a methodology to combat this fraud, subsidizing the work of the auditors and providing greater efficiency to the inspection process. First the state of the art was mapped, identifying which are the strategies used by the scientific community so far and identifying the most relevant works and authors in this research line. Once defined the methodology to be applied, proposed in Silva (2011) and altered by this work, a theoretical framework was developed involving all the concepts of Database Knowledge Discovery (KDD) necessary for the understanding of this study. Done that, the method was described and tested with a study of case applied to a real bidding database, available on the website of *Controladoria Geral da União* (CGU), and the results were promising.

Keywords: Data Mining. KDD. Bidding. Cartels.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CNPJ	Cadastro Nacional de Pessoas Jurídicas
CGU	Controladoria Geral da União
DCBD	Descoberta de Conhecimento em Base de Dados
DM	Data Mining
FBI	Federal Bureau of Investigation
IBPAD	Instituto Brasileiro de Pesquisa e Análise de Dados
ID	Número de Identificação
MIT	Massachusetts Institute of Technology
KDD	Knowledge Discovery in Data Base
MD	Mineração de Dados
ODP	Observatório de Despesas Públicas
PR	Participação Relativa
PRM	Participação Relativa Mínima
RA	Regras de Associação

SUMÁRIO

1 INTRODUÇÃO

A licitação pública foi o instrumento criado pelo legislador de modo perseguir certos princípios implícitos, e também explícitos no artigo 37 da Constituição Federal de 1988, que são: Legalidade, Impessoalidade, Moralidade, Publicidade e Eficiência. Visto que o governo deve pensar no interesse coletivo sempre à frente dos individuais, a licitação tem como objetivo principal tratar os fornecedores interessados de maneira isonômica, e ao mesmo, através de um ambiente competitivo, incentivar a queda dos preços, trazendo benefícios assim à sociedade como um todo (PIETRO, 2019).

Em um processo licitatório, a Legalidade se materializa com a sua previsão em instrumento legal, e o seu cumprimento adequado pelo poder executivo. Já a Impessoalidade é obtida com o tratamento isonômico entre os fornecedores, que são colocados frente a um ambiente competitivo sem vantagens uns com os outros. A Moralidade, que apesar de ter um conceito considerado “impreciso” por muitos doutrinadores, fica evidente quando o processo é conduzido com honestidade perante a sociedade e os participantes do certame. A Publicidade é atingida com a divulgação do trâmite pelos meios adequados e previstos em lei, e é um princípio necessário ao funcionamento eficaz deste instrumento. Por fim, a Eficiência é alcançada através do ambiente competitivo produzido pela licitação, que incentiva a queda dos preços pagos pelo governo, minimizando os gastos e maximizando os benefícios trazidos à sociedade (NIEBUHR; NIEBUHR, 2018).

1.1 O PROBLEMA

Apesar das licitações funcionarem perfeitamente na teoria, há uma série de atos fraudulentos que frustram com os objetivos governamentais citados no tópico anterior, e precisam ser combatidos. Como explicam Santos e Souza (2018), dentre as fraudes mais comuns e onerosas ao Estado estão os cartéis ou rodízios, que podem ser definidos como: Um acordo prévio de preços entre os fornecedores participantes da licitação, que geralmente se revezam na posição vencedora, e eliminam o ambiente competitivo. Não havendo essa competição, os preços se elevam e o governo arca com este prejuízo, enquanto as empresas fraudulentas se enriquecem ilicitamente.

Como descrito no tópico 2, Artigo A, os cartéis podem ocorrer mesmo sem a participação ilícita de um agente público, o que indica que apenas auditorias internas não são capazes de detectá-la. Por essa razão, órgãos de controle externo e estudiosos do mundo todo se dedicam a desenvolver maneiras de combater este conluio (SANTOS; SOUZA, 2018).

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

Propor e testar uma metodologia capaz de detectar a formação de cartéis em licitações públicas. Tal abordagem pretende utilizar Ciência de Dados para tornar mais eficiente os processos de auditoria governamental, aumentando assim a taxa de detecção dessa fraude, e consequentemente diminuindo os prejuízos ao erário.

1.2.2 OBJETIVOS ESPECÍFICOS

- Pesquisar o atual “estado da arte”, de modo a entender até que ponto os pesquisadores do assunto já avançaram, e quais foram os obstáculos e soluções por eles encontradas até o momento. Nesta etapa é essencial conhecer as principais obras no assunto, os principais autores, as metodologias mais relevantes, e as ferramentas utilizadas no combate aos cartéis;
- Propor uma nova metodologia que englobe os avanços obtidos pelos pesquisadores anteriores, visando contribuir para a expansão das fronteiras do conhecimento neste assunto. Como será apresentado mais adiante nesse trabalho, o método proposto utiliza da Descoberta de Conhecimento em Base de Dados (DCBD), onde a etapa de Mineração de Dados (MD) é aplicada por intermédio das Regras de Associação (RA).
- Uma vez definida a metodologia, fazer um estudo de todo o referencial teórico pertinente à mesma, incluindo os conceitos básicos da DCBD, suas mais diversas aplicações, o funcionamento dos algoritmos utilizados, entre outros.
- Realizar um estudo de caso que vise testar e possivelmente validar a metodologia proposta. Para tanto, é necessário coletar uma base de dados aberta sobre processos licitatórios e realizar as cinco etapas da DCBD:

Seleção, Pré-Processamento, Formatação, Mineração de Dados (MD) e Interpretação dos Resultados.

- Analisar os resultados obtidos e avaliar de maneira crítica a metodologia proposta. Identificar seus pontos positivos, capazes de revelar indícios da formação de carteis, e seus possíveis pontos de melhoria futura.

1.3 DELIMITAÇÃO DA PESQUISA

A limitação dessa pesquisa encontra-se na base de dados utilizada para teste e validação da metodologia proposta. Como se sabe, empresas que praticam fraudes de formação de cartel atuam em licitações tanto no âmbito federal, como no estadual e no municipal, não se limitando aos editais lançados por apenas um dos entes em específico. Em virtude disso, para que não houvesse risco de perda de informações, seria necessário analisar uma base de dados integrada, contendo os processos licitatórios das três esferas de governo, entretanto esse recurso ainda não está disponível.

Como será explicado no tópico 4, Artigo C, os dados utilizados neste estudo foram obtidos no portal da Controladoria Geral da União (CGU), portanto, se referem apenas a este ente federativo específico. Entretanto, há dois pontos positivos: O primeiro é que a metodologia proposta continua válida para estudos futuros, caso haja a oportunidade de sua realização em uma base integrada, o que ainda não é realidade no Brasil; O segundo é que, a União certamente é a esfera de governo com mais volume de licitações, logo, dentre as possíveis bases de teste do método, esta certamente é a mais completa e adequada para tanto.

1.4 ESTRUTURA DO TRABALHO

Esta dissertação é composta por cinco capítulos, sendo o primeiro uma introdução geral sobre o tema, e o último as considerações finais acerca do estudo. Já nos capítulos 2, 3 e 4, constam três artigos independentes chamados de Artigo A, Artigo B e Artigo C, que apresentam suas próprias Introduções, Referenciais Teóricos, e Conclusões, porém, cada um com objetivos distintos, de modo a somar para compor o resultado final desse trabalho.

No Artigo A, tópico 2, encontra-se um estudo bibliométrico a fim de se avaliar o “estado da arte”, compreender os principais trabalhos correlatos ao tema, identificar os autores mais importantes, bem como as metodologias utilizadas para solução do problema proposto. Em posse de toda essa informação, foi possível realizar em paralelo o referencial teórico, abrangendo todos os conceitos necessários ao entendimento dessa dissertação, inclusive sobre o método a ser proposto no artigo seguinte.

No Artigo B, tópico 3, se expõe a metodologia proposta neste trabalho, trazendo em riqueza de detalhes o passo a passo para sua execução. Já o teste e validação deste método encontram-se no Artigo C, presente no tópico 4, que comprovou a eficiência e a eficácia da abordagem sugerida nessa dissertação em um estudo de caso real.

A figura 1 apresenta uma síntese do que foi descrito acima, com os tópicos enumerados de 1 a 5, e as respectivas ideias centrais de cada artigo que compõe o trabalho.

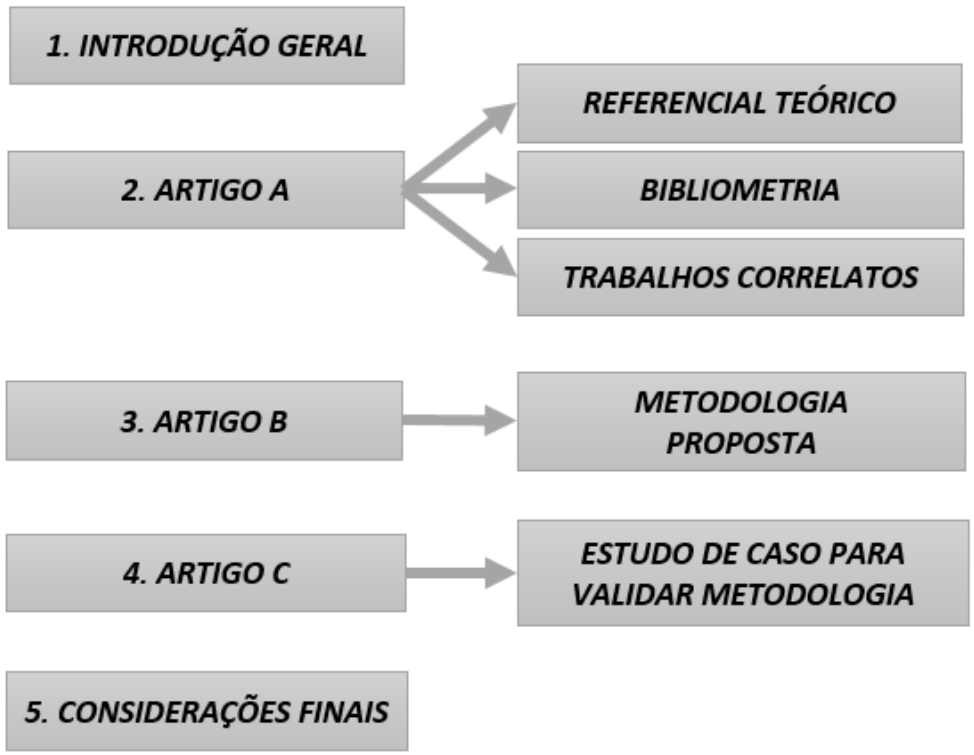

Figura 1 - Estrutura do Trabalho

Fonte: Próprio Autor.

2 ARTIGO A - MINERAÇÃO DE DADOS NO COMBATE AOS CARTÉIS: REFERENCIAL TEÓRICO, BIBLIOMETRIA E TRABALHOS CORRELATOS

Resumo

O objetivo deste artigo é trazer o referencial teórico necessário ao entendimento do estudo da formação de cartéis em processos licitatórios por meio da Mineração de Dados (MD). Para tanto, este trabalho apresenta conceitos relativos à Licitação, instrumento pelo qual a administração pública realiza suas compras, além de todo arcabouço teórico sobre Descoberta de Conhecimento em Base de Dados (DCBD), que tem a Mineração de Dados (MD) como sua principal etapa. Além disso, este artigo também se propõe em realizar uma análise bibliométrica, mapeando os trabalhos correlatos, os principais autores, e as publicações mais relevantes, a fim de se entender o atual “estado da arte”, com vistas a produzir trabalhos futuros que promovam alguma contribuição.

Palavras-chave: Mineração de Dados. DCBD. Licitações. Cartéis.

Abstract

The purpose of this paper is to describe the necessary theoretical framework to understand the study of cartel formation in bidding processes using Data Mining (DM). Therefore, this paper presents concepts related to Bidding, the instrument used by the public administration to do purchases, in addition to the theoretical framework regard Knowledge Discovery in Database (KDD), which has Data Mining (DM) as your main step. In addition, this article also proposes to perform a bibliometric analysis, mapping the related studies, the main authors, and the most relevant publications, making possible to understand the current “state of art”, in order to producing future works with some contribution.

Keywords: Data Mining. KDD. Bidding. Cartels.

2.1 INTRODUÇÃO

Para que a administração pública realize suas compras, ressalvadas as exceções previstas em lei, deve realizar um processo de competição transparente e justo aos interessados, chamado de licitação. Entretanto, há um tipo de fraude muito comum entre os fornecedores, que se organizam em forma de conluio, fazendo um acordo prévio de preços, e eliminando assim a competição. Este tipo de fraude é chamado de cartel, e tem sido um grande desafio para os órgãos responsáveis pela fiscalização dos gastos públicos (SANTOS; SOUZA, 2018).

A administração pública e seus órgãos de controle, como a Controladoria Geral da União (CGU), têm se preocupado cada vez mais com a utilização de tecnologia de ponta nas auditorias governamentais, de forma trazer mais eficiência ao processo de fiscalização. Neste sentido, a CGU criou uma unidade permanente chamada Observatório da Despesa Pública (ODP), que aplica metodologia científica, dentre elas a Mineração de Dados (MD), para subsidiar a tomada de decisões estratégicas relacionadas ao monitoramento dos gastos públicos (CGU, 2019a).

A unidade da ODP tem se mostrado tão eficiente em promover transparência, que desde a sua criação em 2008, já recebeu diversas premiações nacionais e internacionais, como listado a seguir (CGU, 2019a):

- Prêmio TI & Governo (2009);
- Prêmio Conip de Excelência em Inovação na Gestão Pública (2009);
- Prêmio Excelência em Governo Eletrônico (2010);
- United Nations Public Service Awards (2011);
- Prêmio Conip de Excelência em Inovação na Gestão Pública (2013).

Como destacou o cientista de dados chefe da ODP, Rommel N. Carvalho, em seminário realizado pelo Instituto Brasileiro de Pesquisa e Análise de Dados (IBPAD), no dia 15 de junho de 2016, a Mineração de Dados (MD) é uma grande aliada da fiscalização governamental, sendo extremamente útil no combate à corrupção. Nesta mesma ocasião, Rommel apresentou diversos trabalhos de gestão de *Big Data* realizados pelo órgão federal, dentre eles a identificação de riscos de fraudes em licitações com o uso do *Data Mining* (IBPAD, 2016).

Para reforçar ainda mais a importância da MD na fiscalização das licitações públicas, cabe destacar a participação da ODP em congressos com essa temática, a exemplo da *Conference on Knowledge Discovery and Data Mining*, realizada em agosto de 2016, em São Francisco, na Califórnia. Esse evento reuniu palestrantes de universidades renomadas como o Instituto de Tecnologia de *Massachusetts* (MIT) e a Universidade de *Harvard*, além de grandes empresas, como o *Google*, *Facebook*, *Uber* e *Netflix*, todos presentes com um único intuito: trocar informações sobre essa ferramenta poderosa e promissora, que é a Mineração de Dados (CGU, 2016).

2.2 LICITAÇÕES PÚBLICAS

A administração pública é regida, inclusive no que tange às licitações, por princípios expressos no artigo 37 da Constituição Federal de 1998, e outros princípios implícitos em seu texto. A Legalidade, por exemplo, vincula o processo licitatório ao instrumento legal que o previu, fazendo com que as autoridades obedeçam à legislação na realização desses procedimentos. (PIETRO, 2019).

Já a Impessoalidade, aponta no sentido de que o governo deve tratar os indivíduos semelhantes de maneira semelhante, e os diferentes de maneira diferente, na medida de suas diferenças. Com base nessa afirmação, a licitação deve proporcionar um ambiente, onde os fornecedores possam competir de maneira justa entre eles (PIETRO, 2019).

A Moralidade por sua vez, é um importante princípio para que a licitação seja conduzida com honestidade e transparência perante a sociedade. Já a Publicidade, exige que o edital seja divulgado nos meios legais adequados, permitindo assim o alcance da licitação a todos os fornecedores interessados (SILVA, 2019).

Por fim, o último princípio expresso é a Eficiência, que pode ser traduzido como a obrigação do Estado em gerir os recursos públicos da maneira mais eficiente possível, evitando desperdícios, e maximizando assim o bem-estar da população. É importante notar que o ambiente competitivo criado pela licitação, naturalmente já reduz os preços, o que vai ao encontro deste princípio (NIEBUHR; NIEBUHR, 2018).

Em vista do que foi exposto, pode-se observar que a licitação é um processo de compras públicas que visa atender a Impessoalidade e a Eficiência, sempre respeitando também a Legalidade, a Moralidade e a Publicidade. (SILVA, 2019).

2.2.1 MODALIDADES DE LICITAÇÃO

Dentre as modalidades de licitação encontram-se: Convite, Tomada de Preços, Concorrência, Leilão, Concurso e Pregão. O tipo de licitação que deve ser adotado pelo órgão responsável é determinado em função do valor do objeto, e também por outros fatores, como as características do bem ou serviço que será adquirido, do tipo de contrato, entre outros (NIEBUHR; NIEBUHR, 2018).

Conforme explica Silva (2019), cada uma das modalidades de licitação têm suas regras específicas de condução, descritas na lei 10.520 de 2002 para a modalidade Pregão, e na lei 8.666 de 1993 para as demais modalidades. Independente do rito legal a ser seguido, o objetivo é o mesmo: Garantir a competição justa e isonômica entre os fornecedores, e perseguir a eficiência, a qualidade e a economicidade no processo de compra.

- **Concorrência:** Modalidade licitatória mais formal e criteriosa, aplicada a maiores vultos e certas situações específicas, como nos contratos de Concessão. Existe uma fase de habilitação, onde os fornecedores interessados devem comprovar que atendem aos requisitos mínimos, e uma fase de classificação e julgamento, onde os envelopes lacrados são entregues a uma comissão que avalia a melhor proposta, seguindo critérios objetivos estabelecidos previamente em lei e edital.
- **Tomada de Preços:** Modalidade um pouco mais simples que a anterior, aplicável a vultos também um pouco menores. Nesta licitação não há fase de habilitação, pois só podem participar os fornecedores que já estiverem previamente cadastrados. No portal online, o administrador observa qual fornecedor inscrito no edital apresenta o menor preço, e este é contratado.
- **Convite:** Modalidade aplicável a vultos ainda menores, onde a administração convida pelo menos três fornecedores a apresentar suas propostas, e escolhe a melhor. Cabe destacar que, para garantir a isonomia, empresas não convidadas também podem participar, desde que estejam cadastradas e informem sua intenção com antecedência mínima de 24 horas.
- **Concurso:** Modalidade bem diferente das demais, pois a administração fixa previamente o valor do prêmio, e o melhor fornecedor, de acordo com critérios definidos em edital, vence o certame. Esta modalidade está associada com a

compra de trabalhos artísticos, técnicos e científicos, sendo assim, seus critérios de apuração do vencedor são dotados de certa subjetividade.

- **Leilão:** Modalidade bem simples, utilizada para alienação de bens. A administração realiza um leilão presencial ou *online*, com lances sucessivos, e vende seus bens pelo maior preço possível.
- **Pregão:** Esta é a modalidade mais usada atualmente, que surgiu de um aprimoramento das modalidades anteriores, agregando eficiência ao processo. Sempre que um objeto ou serviço for de uso comum, ou seja, puder ser definido com objetividade, o pregão pode ser usado, independentemente do valor do objeto. Nesta modalidade os fornecedores fazem lances sucessivos, presenciais ou *online*, e o menor preço firma um contrato de venda com a administração.

2.2.2 CARTEIS E OUTRAS FRAUDES

Apesar de o processo licitatório ser um grande aliado na busca dos objetivos governamentais de Legalidade, Impessoalidade, Moralidade, Publicidade e Eficiência, os órgãos de controle, bem como a administração pública de uma maneira geral, se deparam com o grande desafio de minimizar as fraudes licitatórias. Segundo Santos e Souza (2018), este problema pode ocorrer com o auxílio do agente público contratante ou não, conforme os exemplos listados a seguir:

- **Edital Restritivo:** Consiste em colocar exigências muito específicas, exageradas, combinadas ou desnecessárias no edital da licitação, eliminando os concorrentes que não são capazes de atendê-las, e direcionando assim o processo. Na prática, há diversos exemplos envolvendo essa fraude, tais como: Exigência de capacidade econômica exagerada, de requisitos técnicos muito específicos ou não justificáveis, ou até mesmo de demonstrações contábeis não previstas por lei;
- **Publicidade Precária:** Um dos requisitos essenciais para que uma licitação ocorra de maneira legitima é que ela seja devidamente divulgada pelos meios adequados previstos em lei. Este processo de publicidade permite que todos os fornecedores interessados possam participar do edital, provendo assim uma competição justa. Um exemplo dessa fraude seria a publicação da licitação em

um jornal de pouca relevância, e a convocação direta e informal de determinado grupo, que certamente seria favorecido.

- **Contratação Direta Indevida por Fracionamento de Despesas:** Tipo de fraude cometida com a ajuda de agente público, que ao invés de realizar uma compra de R$30.000, por exemplo, fraciona a mesma em três compras de R$10.000. Dessa forma, com o valor reduzido, é permitida a contratação com dispensa de licitação, facilitando assim que administrador burle a competição, e direcione a compra para determinado fornecedor.
- **Contratação Direta Indevida por Falsa Exclusividade:** Há situações em que é impossível que ocorra a competição, sendo assim a licitação é inexigível, e a contratação é direta. Um exemplo desse fato seria um posto de gasolina sem concorrentes em uma pequena cidade, ou seja, um caso de fornecedor único. Acontece que muitas vezes um particular pode corromper o agente público a simular uma situação de exclusividade, para que ele seja contratado diretamente sem licitação.
- **Cartel ou Rodízio:** Neste caso não há participação dos agentes públicos, mas sim um acordo prévio de preços entre os fornecedores, que sempre concorrem juntos nas mesmas licitações, e fazem um rodízio para dividir os contratos de maneira igualitária entre eles. Como exemplo, as empresas “A”, “B” e “C”, podem combinar o preço mínimo de venda ao governo em R$100, eliminando assim a competição, e superfaturando o valor do bem.

2.3 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS (DCBD)

Com os constantes avanços na área computacional, o volume de dados gerados e armazenados pelas organizações é cada vez maior. De acordo com Gantz e Reinsel (2012) o volume de dados digitais aumentou de 166 para 988 *exabytes*, entre 2006 e 2010. Além disso, o custo para se gerar um *gigabyte* irá cair de 2 para 0,2 dólares entre 2012 e 2020, ou seja, uma redução de dez vezes do valor original. Este mesmo estudo ainda vai além, e apresenta uma estimativa de crescimento exponencial para geração de dados digitais, atingindo incríveis 40.000 *exabytes* no ano de 2020, conforme mostra a figura 2.

Basta tomar um exemplo do cotidiano para entender a grande evolução da capacidade de armazenamento dos últimos anos: Os primeiros disquetes, da década

de 1970, tinham a capacidade de armazenamento de 80 Kb, que é cerca de nove trilhões de vezes menor que a capacidade atual de um pen-drive, que armazena um *terabyte*, e é acessível por um baixo custo à grande maioria da população. Além da evolução dos discos removíveis e internos, ainda pode-se citar a tecnologia de armazenamento em nuvem, oferecida gratuitamente por grandes empresas, como *Google, Apple* e *Microsoft* (PIXININE, 2017).

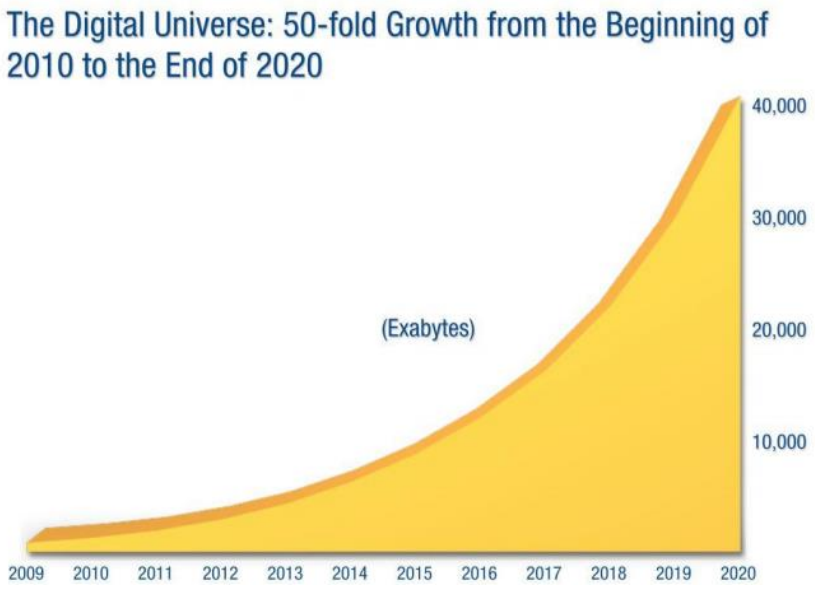

Figura 2 - Crescimento exponencial da geração de dados digitais
Fonte: Gantz e Reisnsel (2012)

Em virtude disso, é cada vez mais barato e viável armazenar *terabytes* de dados, entretanto, isso não significa necessariamente a obtenção de conhecimento. Na verdade, para que o homem consiga obter conclusões úteis em meio a tanta informação, são necessárias ferramentas computacionais adequadas, dentre elas, a Descoberta de Conhecimento em Base de Dados (DCBD). A título de exemplo, este recurso é utilizado atualmente nas mais diversas áreas, tais como: Empresas de comércio eletrônico, tentando entender o perfil dos compradores; empresas do setor financeiro, buscando classificar seus devedores; órgãos governamentais de gestão tributária, visando identificar o perfil dos sonegadores de impostos. De acordo com o jornal americano *The New York Times*, até mesmo a famosa agência de segurança norte americana FBI *(Federal Bureau of Investigation)*, usa a DCBD no combate ao terrorismo, e obtém sucesso (LICHTBLAU, 2007).

O termo DCBD é oriundo do inglês, *Knowledge Discovery in Database* (KDD), e surge na literatura pela primeira vez ao final da década de 1980. Posteriormente é apresentado por Fayyad et al. (1996), já consolidado como um processo sólido, dividido em etapas bem definidas, e com o objetivo de extrair informações úteis e ocultas nos bancos de dados, de modo que possam ser utilizadas para a tomada de decisão, planejamento e gestão.

Feita esta breve introdução, resta agora adentrar no funcionamento deste processo, detalhando o seu passo-a-passo. Como pode ser observado na figura 3, o autor propôs uma estrutura com cinco etapas, capazes de transformar o *input*, que é o Banco de Dados, no *output* desejado, que é o Conhecimento. Em síntese, essas etapas são: Seleção, Pré-Processamento, Formatação, Mineração de Dados (MD) e Interpretação/Avaliação.

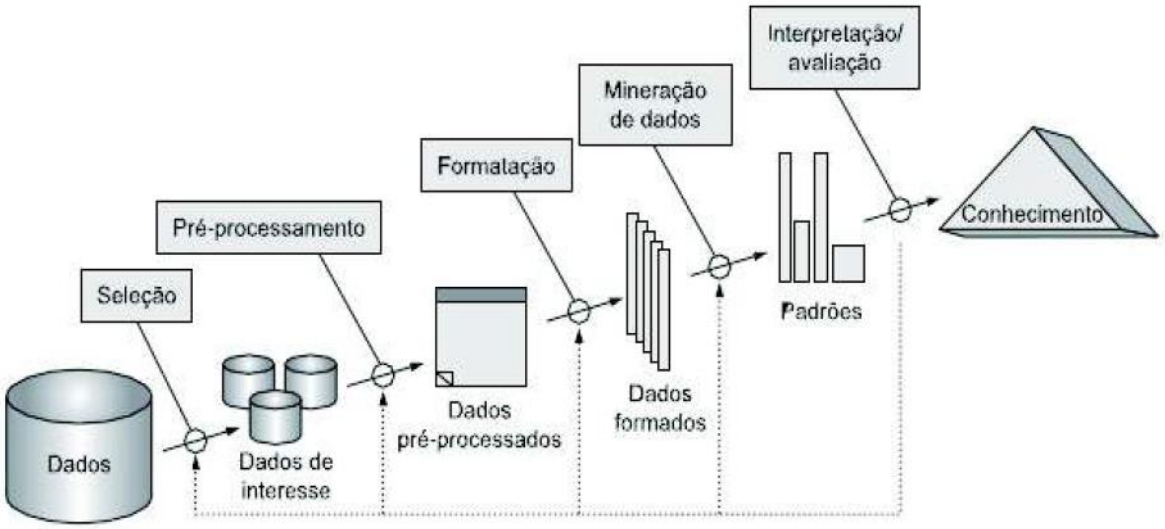

Figura 3 - Input, Etapas e Output da DCBD
Fonte: Adaptada de Fayyad et al. (1996).

2.3.1 ETAPAS DA DCBD

SELEÇÃO

A partir da base de dados original, é preciso selecionar as informações relevantes ao estudo, descartando as demais, que tendem a atrapalhar o processo de mineração e interpretação dos resultados. Para tanto, é necessário conhecer os objetivos da pesquisa, e identificar se determinado atributo pode, ou não, contribuir para geração de conhecimento (FAYYAD et al., 1996).

A título de exemplo, se o objetivo do estudo é identificar a formação de cartéis em processos licitatórios, é nítido que a informação sobre o órgão licitante é irrelevante, já que os cartéis atuam em vários órgãos simultaneamente. Em contrapartida, quais fornecedores participaram de cada licitação é sim uma informação importante, que ajudará a descobrir quando essas empresas atuam juntas em um edital para fraudá-lo.

PRÉ-PROCESSAMENTO

Nessa etapa deve-se fazer a Limpeza de dados inconsistentes, incompletos, duplicados, ou que de alguma forma possam contaminar a base de dados e impedir o funcionamento adequado do método. Além disso, se necessário deve-se proceder a técnica de Integração, que consiste em preenchimento de espaços vazios de maneira estratégica, para posterior mineração (FAYYAD et al., 1996).

A título de exemplo, considerando que uma pequena parte de itens de um banco de dados esteja incompleta, esses poucos registros poderiam ser removidos sem perda de consistência ou volume de informações. Em contrapartida, se muitos itens apresentam um determinado atributo relevante não preenchido, este espaço vazio poderia ser substituído por “não informado”.

FORMATAÇÃO

Como exemplo de Formatação, é possível imaginar um caso em que um dos atributos seja uma variável contínua ou até mesmo discreta, com muitas possibilidades, como a idade dos clientes de uma loja. Caso o cadastro fosse mantido de forma original, o número de diferentes possibilidades seria enorme, certamente com mais de 50 opções de valoração de idades.

Seria então conveniente, agrupar esses dados de modo a reduzir o número de possíveis valores, e ao mesmo tempo dar maior relevância àquela variável transformada. Neste exemplo, uma opção seria: Até 19 anos; de 20 a 39 anos; de 40 a 59 anos; a partir de 60 anos. Esta de Formatação é chamada de Transformação.

Esta etapa deve ser feita de maneira cuidadosa, pois a redução das possíveis variáveis não pode implicar na perda das propriedades de determinado grupo. Um possível exemplo de formatação com perda de propriedade poderia ser: Até 49 anos; a partir de 50 anos. Nessa situação, muita informação relevante poderia se perder, não sendo uma Transformação bem aplicada.

MINERAÇÃO DE DADOS

A Mineração de Dados (MD) é a etapa mais importante do processo. Sua relevância é tamanha, que frequentemente é confundida e tratada como sinônimo do DCDB. Cabe salientar que, enquanto esse é o processo como um todo, aquela é apenas uma de suas etapas.

É importante observar que a Seleção, o Pré-processamento e a Formatação serviram de fases auxiliares à MD, no sentido de preparar os dados para que pudessem ser minerados adequadamente. Muitas são as tarefas e técnicas para se extrair padrões, tendências e conhecimento de um banco de dados, sendo assim, este artigo reserva o capítulo seguinte para comentar esta etapa com mais riqueza de detalhes (FAYYAD et al., 1996).

INTERPRETAÇÃO

É a última etapa da DCBD, destinada a analisar os resultados obtidos com a MD, a fim de se reconhecer informações úteis de acordo com o objetivo da pesquisa. Em virtude disso, é importante que haja neste caso a participação de entendedores do negócio ou do objeto. Cabe dizer ainda, que como a DCDB é um processo interativo e iterativo, este é o momento de avaliar as etapas anteriores, fazer alterações, e verificar seus efeitos nos resultados (FAYYAD et al., 1996).

A título de exemplo, é possível fazer testes incluindo ou excluindo determinados atributos, o que seria uma alteração na etapa de Seleção. Outro exemplo seria variar os critérios adotados para realizar o agrupamento das variáveis em intervalos, como explicado na etapa de Formatação. Cabe ressaltar, que tais alterações influenciam nos resultados, cabendo ao autor julgar as etapas anteriores a partir dos resultados obtidos.

2.3.2 MINERAÇÃO DE DADOS

Como mencionado no tópico anterior, a Mineração de Dados (MD) ou *Data Mining* (DM), é a etapa em que definitivamente serão processados os dados, a fim de se obter padrões ocultos e informações úteis. No entanto, o termo MD é amplo, e engloba diversas Tarefas distintas, que por sua vez produzem resultados distintos, de acordo com seus objetivos. Na tabela 1, adaptada de Camilo e Silva (2009), pode-se observar um resumo das principais Tarefas da MD.

Tabela 1 - Principais Tarefas da MD

TAREFA	CARACTERÍSTICAS	EXEMPLO DE APLICAÇÃO
Descrição	Utilizada para descrever padrões e tendências em dados.	Geralmente utilizada em conjunto com técnicas de análise exploratória de dados;
Classificação	Uma das tarefas mais comuns da MD, busca associar um conjunto de atributos à um atributo objetivo, categórico, chamado de atributo chave.	Determinar quando uma transação de cartão de crédito pode ser uma fraude;
Estimativa ou Regressão	Tarefa similar à classificação, porém, é utilizada quando o conjunto de atributos em questão é identificado por um valor numérico e não categórico.	Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal;
Predição	Tarefa similar às tarefas de classificação e regressão, porém, esta visa prever o valor de um terminado atributo.	Predizer o valor de uma ação três meses adiante;
Agrupamento (Clusterização)	Tarefa que não tem atribuição de classificação, regressão ou predição, mas sim a de agrupar as instâncias dos dados conforme os valores de seus atributos, não necessitando a definição de um atributo alvo, ou seja, as instâncias não são categorizadas, como na classificação.	Para auditoria, separando comportamentos suspeitos;
Associação	Busca identificar a relação entre os atributos, apresentando-se na forma SE ocorre A, ENTÃO ocorre B.	Identificar os usuários de planos que respondem bem a oferta de novos serviços;

Fonte: Adaptado de Camilo e Silva (2009).

Além disso, cabe dizer que para a realização de cada Tarefa, existem ainda diferentes Técnicas de mineração, que são formas ou métodos para se atingir determinado objetivo. A tabela 2, também adaptada de Camilo e Silva (2009), resume as principais Técnicas de MD.

Tabela 2 - Principais Técnicas de MD

MÉTODO	CARACTERÍSTICAS
Árvores de Decisão	Fluxograma top-down em forma de árvore onde cada nó (atributo) indica um teste a ser feito sobre um valor. Cada nó inferior está ligado e representa um possível valor do nó superior. As folhas indicam qual classe a instância pertence. A partir da estrutura da árvore, extraem-se as regras.
SVM	Técnica permite gerar modelos lineares e não-lineares, podem ser utilizadas para tarefas de classificação e predição.
Classificação Bayesiana	Técnica estatística baseada no teorema de Thomas Bayes que diz ser possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu: Probabilidade (B dado A) = Probabilidade(A e B)/Probabilidade(A). O algoritmo considera a inexistência de relação de dependência entre os atributos que, nem sempre isto é possível.
Redes Neurais	Com origem na psicologia e na neurobiologia, esta técnica simula o comportamento dos neurônios humanos. A rede possui um conjunto de entradas, às quais são aplicados pesos gerando saídas. Ao logo do processo de aprendizado, os pesos são ajustados a fim de aumentar a taxa de acerto de classificações corretas.
Algoritmo Genético	Seguindo a teoria da evolução onde o mais forte prevalece, o algoritmo, a partir de um estado inicial, passa por inúmeras iterações, simulando a seleção natural pelas melhores soluções.

Fonte: Adaptado de Camilo e Silva (2009).

Como este trabalho utilizou a Tarefa e Associação e o algoritmo Apriori no estudo da formação de cartéis em licitações, o tópico subsequente descreverá em detalhes as características desta Tarefa, e o funcionamento deste algoritmo, que cria as chamadas Regras de Associação (RA).

2.3.3 ASSOCIAÇÃO E ALGORITMO APRIORI

A Associação tem o objetivo de identificar quando fatos ocorrerem de maneira correlacionada ou simultânea. Trata-se de uma relação de causa e consequência, do tipo “Se... então...”. Um clássico exemplo sobre a Tarefa de *Associação* pode ser encontrado em Larose (2005), que descreve um estudo feito em um banco de dados de um supermercado. O referido trabalho apresentou que, em determinado dia e horário da semana, SE os clientes compravam fraldas, ENTÃO também compraram cerveja, o que foi uma constatação completamente inesperada.

Cabe ressaltar que esse é o objetivo da DCBD: Tirar conclusões ocultas, não óbvias, escondidas no banco de dados. Seria fácil perceber que o cliente que compra pão, também compra manteiga ou leite, mas a tarefa de Associação foi muito além. Ao descobrir que as clientes compram fralda e cerveja em conjunto, o supermercado teve posse de um conhecimento novo, e pôde criar estratégias de venda orientadas por essa constatação.

Entendido em linhas gerais o funcionamento dessa Tarefa de MD, é necessário entender sobre a Técnica de geração dessas Regras de Associação (RA). Para ilustrar o problema, vamos supor a base de dados apresentada em Silva (2011), constante na tabela 3, com 14 registros que abrangem os seguintes atributos: Tempo, Temperatura, Umidade, Ventando e Jogar.

Tabela 3 - Exemplo de Base de Dados

	Tempo	Temperatura	Umidade	Ventando	Jogar?
1	ensolarado	quente	alta	falso	não
2	ensolarado	quente	alta	verdadeiro	não
3	nuvens	quente	alta	falso	sim
4	chuvoso	moderado	alta	falso	sim
5	chuvoso	frio	normal	falso	sim
6	chuvoso	frio	normal	verdadeiro	não
7	nuvens	frio	normal	verdadeiro	sim
8	ensolarado	moderado	alta	falso	não
9	ensolarado	frio	normal	falso	sim
10	chuvoso	moderado	normal	falso	não
11	ensolarado	moderado	normal	verdadeiro	sim
12	nuvens	moderado	alta	verdadeiro	sim
13	nuvens	quente	normal	falso	sim
14	chuvoso	moderado	alta	verdadeiro	não

Fonte: Silva (2011).

Supondo a regra a seguir, é possível calcular suas medidas de qualidade: Suporte e Confiança. Nesta situação, o lado esquerdo da regra (Temperatura = Frio, Umidade = Normal) aparece nas linhas 5, 6, 7 e 9, ou seja, se repete em 28% dos registros, e esse é o seu Suporte. Observe agora que, dentre esses quatro registros, o lado direito da regra (Jogar = Sim) é atendido três vezes, ou 75%, que é a sua medida de Confiança.

Temperatura = Frio; Umidade = Normal → Jogar = Sim.
Suporte = 28%; Confiança = 75%

Resta então detalhar o funcionamento do algoritmo capaz de criar tais regras, baseado em valores mínimos de Suporte e Confiança. O primeiro desafio para criação deste algoritmo foi: Para base de dados maiores, é impossível calcular exaustivamente o Suporte e a Confiança de todas as regras possíveis em tempo hábil, já que a complexidade deste problema é de: $3^d - 2^d + 1$, onde "d" é o número de itens da base de dados (TAN; STEINBACH; KUMAR, 2009).

Com intuito de resolver esse problema, foi proposto em Agrawal e Srikant (1994) o algoritmo de Associação mais influente e usado até hoje, chamado de Apriori. Este algoritmo atua de duas formas: a poda por Suporte Mínimo, e a poda por Confiança Mínima. Essas duas estratégias encontram-se detalhadas em Tan, Steinbach e Kumar (2009) e Silva (2011), e serão explicadas a seguir:

PODA BASEADA NO SUPORTE:

Para entendimento deste método é necessário compreender previamente o seguinte princípio: "Se um conjunto é frequente, então todos os seus subconjuntos também serão frequentes".

Ou seja, se o conjunto {a, b, c} aparece frequentemente nos registros de dados, logo, todos os seus subconjuntos {a, b}, {a, c}, {b, c}, {a}, {b}, {c} também terão a mesma frequência, ou maior. Analogamente, se um subconjunto {a, b} não é frequente, então seus superconjuntos {a, b,...} também não serão.

Entendido este princípio, pode-se, a partir de um Suporte Mínimo (frequência de aparição na base), eliminar todos os subconjuntos e seus superconjuntos que não atendem o requisito desejado. É importante notar que esta poda reduz consideravelmente a complexidade do problema, permitindo assim a criação das Regras de Associação (RA) em bases de dados volumosas.

PODA BASEADA NA CONFIANÇA:

Nessa abordagem, primeiramente o algoritmo gera regras cujo consequente contenha apenas um item, como exemplo: {a, c, d} → {b} e {a, b, d} → {c}. Caso elas apresentem Confiança alta, o algoritmo faz a fusão das mesmas, gerando uma nova regra: {a, d} → {b, c}. A nova Confiança é calculada, e se ela atende ao mínimo exigido, deve ser mantida.

Analogamente, se uma regra gerada inicialmente com apenas um item no seu subsequente possui Confiança baixa, então todas as regras contendo aquele mesmo consequente também são descartadas, reduzindo assim consideravelmente a complexidade do problema. Em outras palavras, se {b, c, d} → {a} é uma regra de baixa Confiança, então {c, d} → {a, b} e {b, d} → {a, c} também são, logo, devem ser eliminadas.

2.4 BIBLIOMETRIA

Antes de iniciar qualquer pesquisa científica, é essencial entender em que contexto ela se encontra e os avanços que já foram obtidos por trabalhos anteriores, o que é comumente chamado de “estado da arte”. Para tanto, algumas perguntas devem ser respondidas:

- Quem são os pesquisadores relevantes neste tema?
- O que já foi descoberto por estudos anteriores?
- Ainda há algo de novo com que este trabalho possa contribuir?

Respondendo a essas perguntas antes de se iniciar um novo estudo, o pesquisador adquire “bagagem” sobre o tema, e desta forma é capaz de orientar de maneira mais racional seus esforços, aumentando assim suas chances de contribuir com um conhecimento novo. Em vista disso, este tópico traz um estudo bibliométrico acerca do tema.

Foi realizada no dia 02 de setembro de 2019, uma pesquisa na base de dados *Scopus Elsevier* acerca desta temática. Observe na figura 4 que o número de publicações que contém o termo “MINERAÇÃO DE DADOS”, ou “INTELIGÊNCIA ARTIFICIAL”, ou “*DATA MINING*”, ou “KDD”, “*ARTIFICIAL INTELLIGENCE*”, ou “*BIG DATA*”, em seu título, resumo ou palavras-chave, chega a mais de 522 mil. Nas

mesmas condições, se a pesquisa for realizada pelos termos “LICITAÇÕES”, ou “COMPRAS PÚBLICAS”, ou “*BIDDING*”, ou “*PUBLIC PURCHASES*”, o resultado atinge mais de 15 mil publicações.

Ainda na figura 4, note que os documentos que falam ao mesmo tempo sobre os dois temas supracitados atingem cerca de 600 publicações, e estas serão alvos da uma análise bibliométrica, a fim de se extrair informações relevantes.

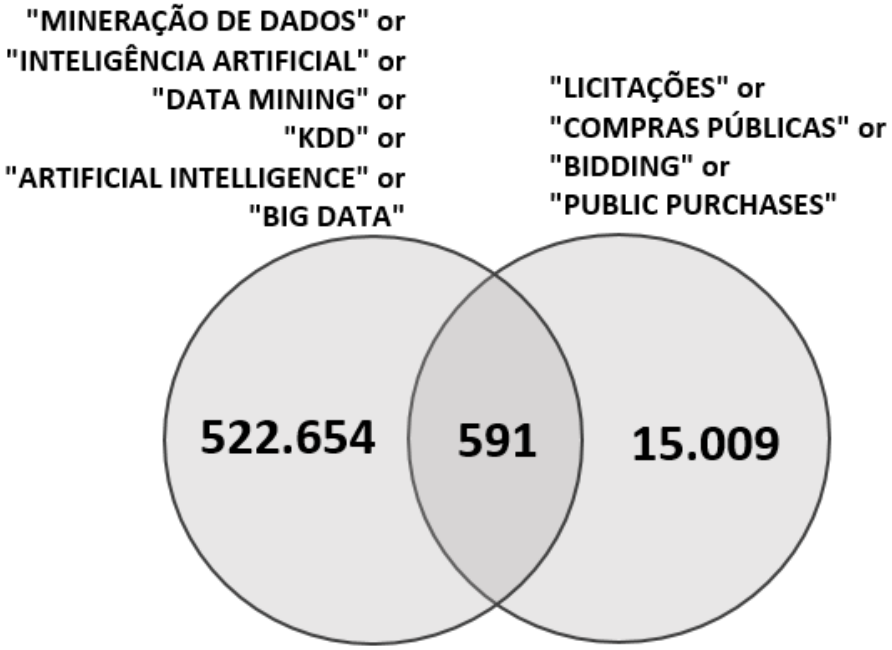

Figura 4 - Número de publicações sobre o tema

Fonte: Próprio Autor.

Com intuito de medir o interesse da comunidade científica, foi feita uma análise para verificar o número de publicações anuais sobre o uso de Mineração de Dados (MD) ou Inteligência Artificial aplicada a licitações públicas. Na figura 5 pode-se observar uma crescente evolução no número de publicações, desde o primeiro registro em 1979, até a data desta pesquisa, em 2019. Já na figura 6, foi feito um recorte dos últimos dez anos de pesquisa, e se observou que o número de publicações tem oscilado, com uma leve tendência de crescimento.

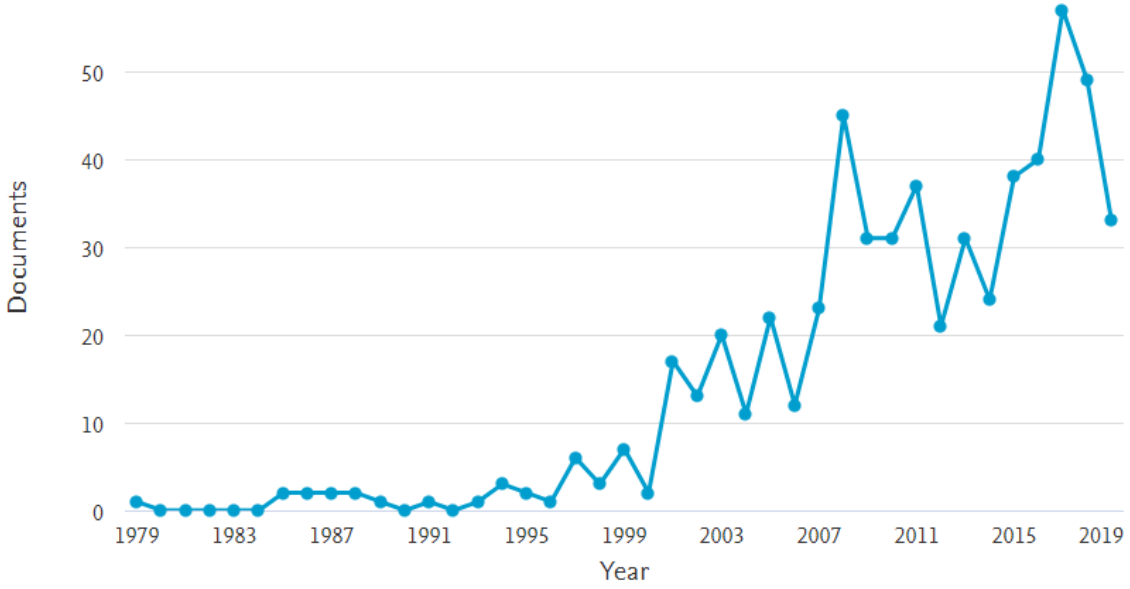

Figura 5 - Evolução anual das publicações desde o primeiro registro

Fonte: Próprio Autor.

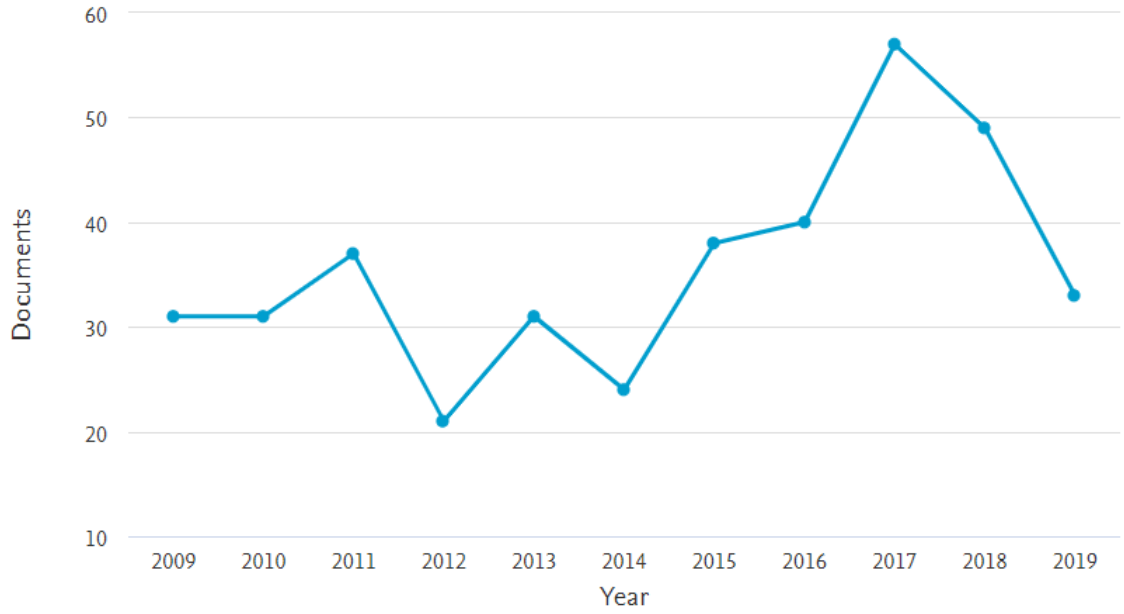

Figura 6 - Evolução anual das publicações nos últimos dez anos

Fonte: Próprio Autor.

Uma importantíssima análise a ser feita é conhecer dos autores que mais se destacam nesta linha de pesquisa, a fim de se fazer uma leitura de seus trabalhos. Observe na figura 7, que o pesquisador chinês *Yong Yuan*, filiado na *Qingdao Academy of Intelligent Industries*, lidera o ranking com 17 publicações, seguido por *Nicholas R. Jennings*, filiado na *Imperial College London*, com 15 registros.

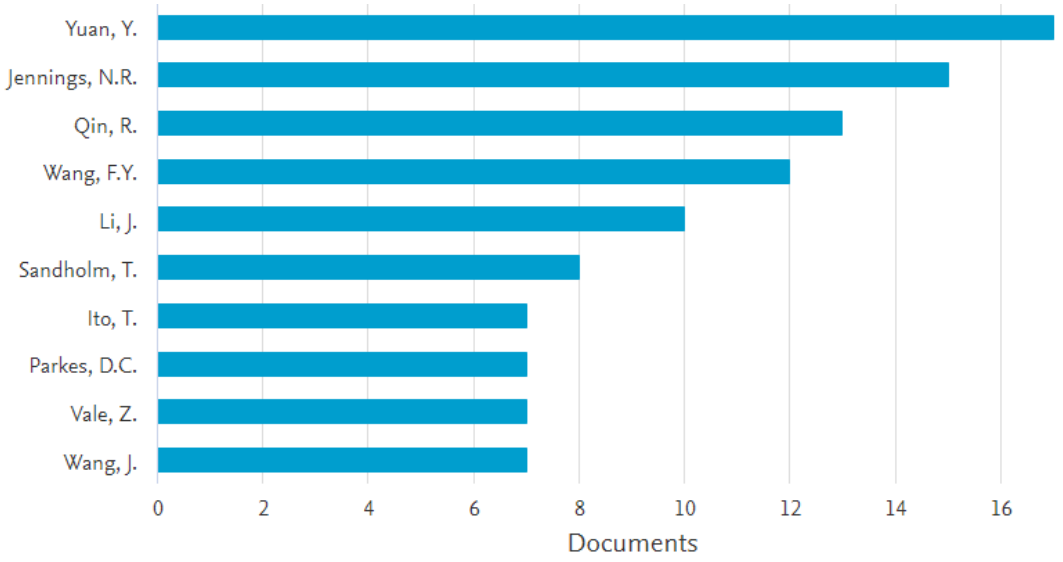

Figura 7 - Autores que mais publicaram sobre o tema

Fonte: Próprio Autor.

Outra análise que pode ser feita é sobre o número de publicações de cada país. Na figura 8 observam-se as dez nações que mais publicam sobre Ciência de Dados aplicada às licitações, e o Brasil, que ocupa a 33º posição. É notório que, enquanto o líder Estados Unidos tem mais de 160 registros, seguido pela China, com mais de 120, o Brasil apresenta apenas duas publicações. Isso sugere que, apesar de este ser um assunto pesquisado por estudiosos do mundo todo, o tema não é muito explorado no Brasil.

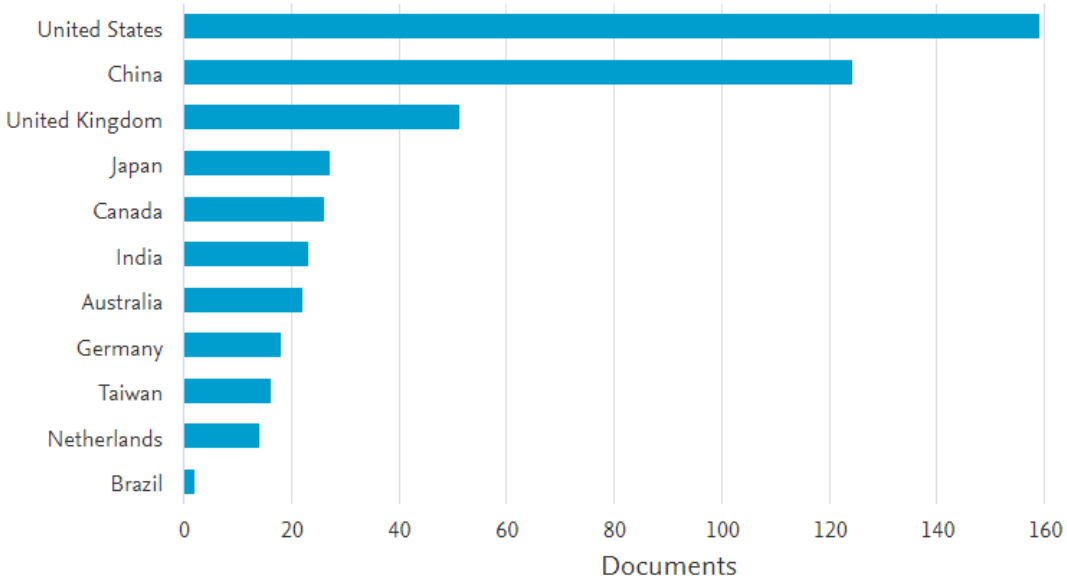

Figura 8 - Número de publicações por país

Fonte: Próprio Autor.

Ainda na base *Scopus Elsevier*, pode-se perceber quais são as áreas que mais se relacionam com o assunto. Como apresentado na figura 9, Ciência da Computação domina a linha de pesquisa, com 42,4% das publicações, seguida por Engenharia e Matemática. Isso é compreensível, que já Mineração de Dados (MD) envolve o uso de algoritmos, bases de dados computacionais, manuseio de softwares, e outras ferramentas afins.

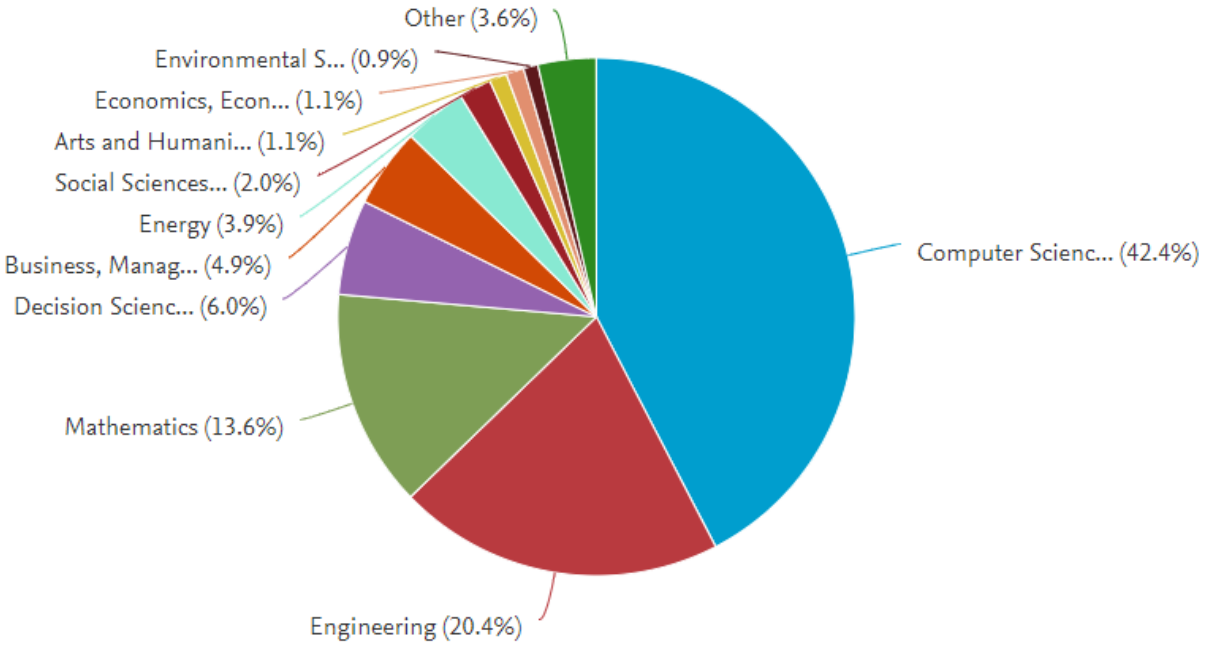

Figura 9 - Número de publicações por área

Fonte: Próprio Autor.

A última análise realizada para se buscar trabalhos correlatos foi a de verificar as universidades que mais publicam sobre o estudo das licitações por meio de Mineração de Dados (MD). Pode-se observar na figura 10, que a *Chinese Academy of Sciences,* localizada na China, lidera esse *ranking* com 19 publicações. Além disso, a *University of Southampton*, na Inglaterra, também possui relevância no assunto, com 18 registros nesta linha de pesquisa.

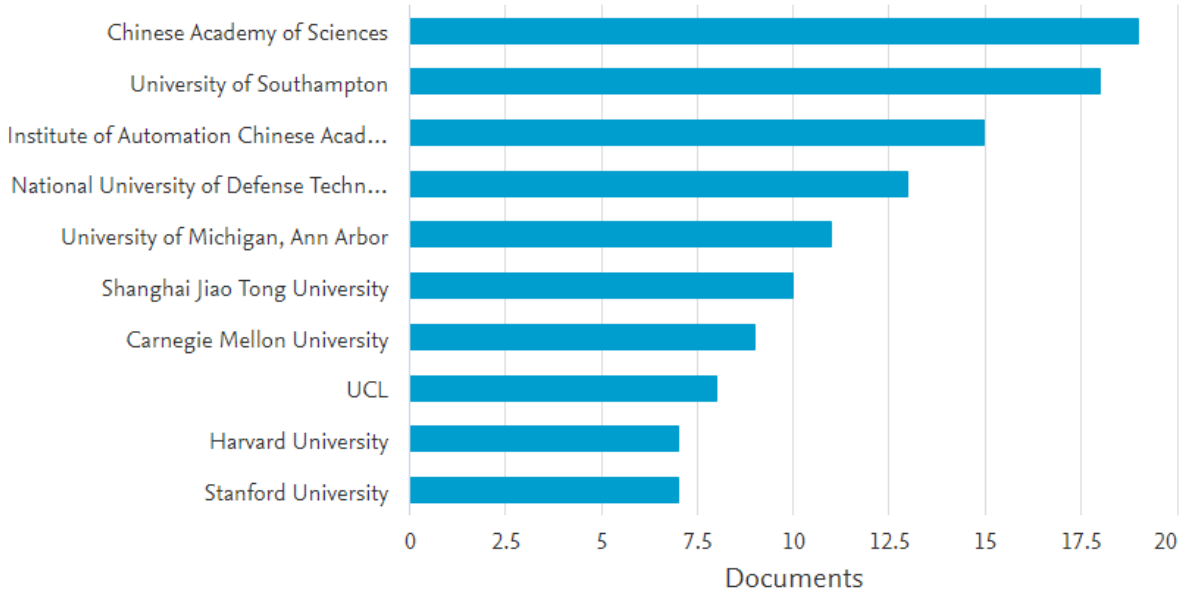

Figura 10 - Número de publicações por universidade

Fonte: Próprio Autor.

Com base em todas as análises supracitadas, foi possível identificar quais são as tendências atuais que relacionam o estudo das licitações públicas por meio da Ciência de Dados, e assim direcionar esta pesquisa.

2.4.1 REFINAMENTO DA PESQUISA

Outra estratégia adotada a fim de se encontrar trabalhos correlatos foi a de limitar o universo da pesquisa anterior, somente àqueles trabalhos que versam sobre a formação de cartéis. Com essa finalidade, uma nova *Query* foi criada da seguinte maneira: (“MINERAÇÃO DE DADOS” ou “INTELIGÊNCIA ARTIFICIAL” ou “*DATA MINING*” ou “KDD” ou “*ARTIFICIAL INTELLIGENCE*” ou “*BIG DATA*”) e (“LICITAÇÕES” ou “COMPRAS PÚBLICAS” ou “*BIDDING*” ou “*PUBLIC PURCHASES*”) e (“CARTEL” ou “CARTÉIS” ou “CONLUIO” ou “*CARTELS*” ou “*BID RIGGING*” ou “*COLLUSION*”).

Como resultado, pode-se observar na figura 11 que foram encontradas apenas seis publicações unindo o estudo da formação de cartéis em licitações públicas por meio da Ciência de Dados, em toda a base *Scopus Elsevier*. Essas publicações também foram estudadas a fim de se adquirir bagagem sob o tema, e buscar informações sobre as metodologias adotadas pelos cientistas de dados no combate a essa fraude.

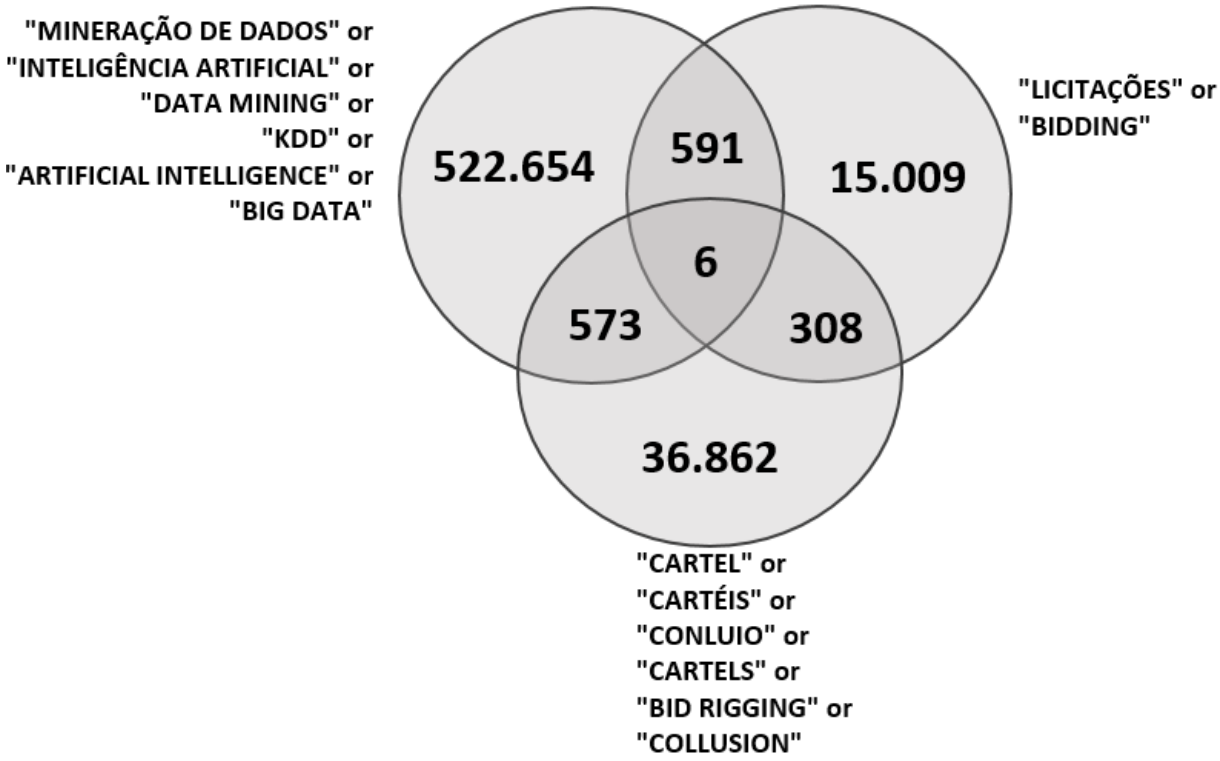

Figura 11 - Número de publicações em pesquisa refinada sobre o tema

Fonte: Próprio autor.

2.5 TRABALHOS CORRELATOS

Dentre os autores lidos, um dos destaques foi o pesquisador Carlos Vinicius Sarmento Silva, graduado em Ciência da Computação pela Universidade de Brasília em 2006, e mestre em Informática pela mesma universidade em 2011. Segundo o site *Google Scholar*, o autor produziu, além de sua dissertação de mestrado, sete publicações entre 2010 e 2012, todas na mesma linha de pesquisa: Detecção de cartéis em processos licitatórios por meio de agentes de Mineração de Dados (MD). Como pode ser visto em Silva (2011), o autor utiliza uma base de dados contendo licitações realizadas pela União, para executar as tarefas de *Clusterização* e Associação, a fim de determinar quais empresas tendem a participar dos certames em conjunto, o que é um forte indicativo da formação de cartéis.

Outro trabalho relevante pode ser encontrado em Gabardo e Lopes (2014), que utilizam da análise de Redes Sociais e Redes Complexas, no intuito de revelar as empresas que formam cartéis para fraudar os processos licitatórios. Além disso, este trabalho avalia se as empresas que agem em conluio apresentam maiores índices de sucesso, ou não, se comparadas às idôneas.

No trabalho de Bajari e Ye (2003), é possível observar uma metodologia capaz de identificar a formação de cartéis pelo comportamento dos lances de uma licitação. Primeiramente foi introduzido um modelo geral de lances não manipulados, ou seja, sem qualquer fraude. Em seguida, foram discutidas as condições necessárias e suficientes para que se possa considerar o comportamento dos lances como competitivos e não manipulados. Feito isso, foi usado o Teorema de Bayes para comparar as distribuições de lances em licitação, a fim de se detectar se o comportamento é normal ou não.

Em recente trabalho, encontrado em Majadi, Trevathan e Bergmann (2018), foi criado um algoritmo que determina em tempo real, se os lances de um leilão estão sendo feitos por livre competição, ou manipulados. Isso pode ocorrer quando o vendedor, que quer elevar o preço de seu produto, age em conluio com um grupo de empresas para que estas ofereçam lances falsos no leilão. Os resultados experimentais do algoritmo testado em um leilão real se mostraram promissores, e podem contribuir para o combate desse tipo de cartel.

O artigo de Sanchez-Graells (2019) traz a avaliação da ferramenta de Inteligência Artificial utilizada pelo Reino Unido no combate à formação de cartéis, chamada de *Screening for Cartels*. Este aplicativo apresenta doze testes com algoritmos distintos, a fim de medir o nível de suspeita de determinada empresa, chamado de *Suspicion Score*. Dentre os testes aplicados, pode-se citar como exemplo a comparação do texto das propostas dos licitantes perdedores: Caso eles sejam muito semelhantes, há um indicativo que foram feitos pela mesma pessoa, o que sugere a formação de um cartel.

Por fim, como último trabalho correlato, pode-se citar Padhi e Mohapatra (2011), que traz uma metodologia capaz de diferenciar os lances fraudulentos dos competitivos em uma licitação. O método é dividido em sete etapas, e utiliza de técnicas estatísticas como média, mediana, variância, teste de assimetria, entre outros, fornecendo um gráfico de controle ao final, que divide os lances em dois *clusters*: Fraudulento ou Competitivo. O algoritmo foi testado em licitações de um departamento de construções na Índia, e se mostrou promissor.

2.6 CONCLUSÃO

O presente artigo teve como um dos objetivos detalhar o referencial teórico sobre a Mineração de Dados (MD) aplicada no combate à formação de cartéis em licitações públicas, mais precisamente com a Tarefa de Associação. Nesta ocasião, foi detalhado o funcionamento do algoritmo mais influente e utilizado para esse fim, chamado Apriori. Além disso, o trabalho também trouxe o referencial teórico sobre licitações públicas, e as fraudes mais comuns contra este processo de compra, a fim de contextualizar o trabalho.

Após o arcabouço teórico, foi feita uma análise bibliométrica sobre este assunto na base *Scopus Elsevier*, atividade esta que foi essencial para compreensão do atual "estado da arte". Foram compreendidos neste ponto, quais são os autores que se destacam nessa linha de pesquisa, suas publicações mais relevantes, quais são universidades estão na vanguarda, além dos países que dominam este *ranking*. Tudo isso serviu para a seleção e leitura de artigos e dissertações correlatas com o tema, permitindo assim o desenvolvimento de trabalhos futuros que possam contribuir, de alguma forma, com o avanço da pesquisa neste tema.

Apesar de a bibliometria ter sido realizada apenas na base *Scopus Elsevier*, esta foi considerada satisfatória, já que esta base é conhecida por possuir um grande acervo de publicações. Além disso, cabe frisar que a busca por trabalhos correlatos não se limitou a esta fonte, sendo encontrados documentos relevantes em outras bases, enriquecendo assim o aprendizado sobre o tema.

2.7 REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. *In:* INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994. **Proceedings** […]. Hove, East Sussex: Morgan Kaufmann, 1994. p. 487-499.

BAJARI, P.; YE, L. **Deciding Between Competition and Collusion.** Review of Economics and Statistics, Massachusetts, v. 85, n. 4, p. 971–989, Massachusetts, nov. 2003.

CAMILO, C.; SILVA, J. **Mineração de Dados:** Conceitos, Tarefas, Métodos e Ferramentas. Goiânia: Universidade Federal de Goiás, 2009.

CGU. **Observatório da Despesa Pública participa de congresso internacional sobre mineração de dados**. 2016. Disponível em: https://www.cgu.gov.br/noticias/2016/08/observatorio-da-despesa-publica-participa-de-congresso-internacional-sobre mineracao-de-dados. Acesso em: 5 ago. 2019.

CGU. **Observatório da Despesa Pública**, 2019a. Disponível em: https://www.cgu.gov.br/assuntos/informacoes-estrategicas/observatorio-da-despesa - publica. Acesso em: 5 ago. 2019.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Advances in knowledge discovery and data mining**. Menlo Park, Calif.: AAAI Press, USA, 1996.

GABARDO, A. C.; LOPES, H. S. Using Social Network Analysis to Unveil Cartels in Public Bids. *In*: EUROPEAN NETWORK INTELLIGENCE CONFERENCE (ENIC), 14., 2014, Wroclaw, Poland. **Anais** […]. Washington: IEEE, 2014. p. 17-21.

GANTZ, J.; REINSEL, D. **The Digital Universe In 2020:** Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. IDC Corporate, Framingham, dez. 2012. Disponível em: https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf. Acesso em: 14 jun. 2019.

IBPAD. **Ciência de Dados no Combate à Corrupção. 2016**. Disponível em: https://www.ibpad.com.br/blog/analise-de-dados/ciencia-de-dados-no-combate corrupcao/. Acesso em: 5 ago. 2019.

LAROSE, D. T. **Discovering knowledge in data:** an introduction to data mining. Hoboken, N.J. : Wiley-Interscience, 2005.

LICHTBLAU, E. **F.B.I. Data Mining Reached Beyond Initial Targets**. The New York Times, Washington, 9 set. 2007. Disponível em: https://www.nytimes.com/2007/09/ 09/washington/09fbi.html. Acesso em: 14 jun. 2019.

MAJADI, N.; TREVATHAN, J.; BERGMANN, N. Real-Time Collusive Shill Bidding Detection in Online Auctions. *In:* MITROVIC, T.; XUE, B.; LI, X. (eds.). ***AI 2018:*** Advances in Artificial Intelligence. Cham: Springer International Publishing, 2018. v. 11320. p. 184–192.

NIEBUHR, J. de M.; NIEBUHR, P. de M. **Licitações e contratos das estatais**. Belo Horizonte: Fórum, 2018.

PADHI, S. S.; MOHAPATRA, P. K. J. Detection of collusion in government procurement auctions. **Journal of Purchasing and Supply Management**, London, v. 17, n. 4, p. 207–221, dez. 2011.

PIETRO, M. S. Z. D. **Direito Administrativo**. 32. ed. São Paulo: Forense, 2019.

PIXININE, J. **Na memória:** relembre a evolução dos dispositivos de armazenamento. 2017. Disponível em: https://www.techtudo.com.br/listas/noticia/ 2015/05/na-memoria-relembre-a-evolucao-dos-dispositivos-de-armazenamento.html. Acesso em: 8 maio. 2019.

SANCHEZ-GRAELLS, A. ‘Screening for Cartels’ in Public Procurement: Cheating at Solitaire to Sell Fool’s Gold? **Journal of European Competition Law & Practice**, Oxford, v. 10, n. 4, p. 199–211, 1 abr. 2019.
SANTOS, F. B.; SOUZA, K. R. de. **Como combater a corrupção em Licitações**. 2. ed. Belo Horizonte: Fórum, 2018.

SILVA, C. V. S. **Agentes de Mineração e sua Aplicação no Domínio da Auditoria Governamental**. Brasília: Universidade de Brasília, 2011.

SILVA, J. A. **Curso de Direito Constitucional Positivo.** 42. ed. São Paulo: Malheiros, 2019.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

3 ARTIGO B - MINERAÇÃO DE DADOS NO COMBATE AOS CARTÉIS: METODOLOGIA PROPOSTA

Resumo

Com fins de garantir a isonomia, moralidade e economicidade, a administração realiza licitações públicas antes de contratar serviços ou comprar bens. Contudo, uma fraude denominada cartel ou rodízio, que é um acordo prévio de preços, elimina a concorrência entre os fornecedores, aumentando assim o preço pago pelo governo. O objetivo deste artigo é propor uma metodologia capaz de auxiliar as auditorias governamentais na detecção deste conluio, por intermédio de Mineração de Dados (MD). Esta abordagem é inspirada em Silva (2011), obtida através de uma simplificação do método original, e uma proposta de modificação no algoritmo que ranqueia as Regras de Associação (RA) obtidas. O estudo de caso aplicando a metodologia proposta neste artigo será apresentado em trabalhos futuros, juntamente com a discussão de resultados e avaliação desta proposta.

Palavras-chave: Mineração de Dados. DCBD. Licitações. Cartéis.

Abstract

In order to guarantee the isonomy, morality and economy, the administration carries out public bids before hiring services or does purchases. However, a fraud called cartel, which is a prior price agreement, eliminates the competition between suppliers and increase the price paid by the government. The purpose of this paper is to propose a methodology capable of assisting government audits in detecting this collusion through Data Mining (DM). This approach is inspired by Silva (2011), obtained by a simplification of the original method, and a proposal to modify the algorithm that ranks the obtained Association Rules. The study of case applying the methodology proposed in this article will be presented in future works, along with the discussion of results and evaluation of this proposal.

Keywords: Data Mining. KDD. Bidding. Cartels.

3.1 INTRODUÇÃO

Antes de realizar a compra ou venda de um bem ou serviço, o governo realiza um processo administrativo chamado de licitação, que visa garantir ao mesmo tempo: A igualdade de competição entre os fornecedores e os melhores preços para os cofres públicos. Entretanto, há um problema que assombra com os objetivos da administração, que é uma fraude denominada cartel ou rodízio (PIETRO, 2019).

Os cartéis são fraudes em que as empresas que participam da licitação combinam previamente o preço, eliminando assim a competição. Geralmente determinado grupo que age em conluio faz o rodizio da empresa vencedora, e as demais oferecem propostas falsas, de modo que a administração acaba por contratar por um preço mais elevado (SANTOS; SOUZA, 2018).

Neste cenário, órgãos de controle governamental de diversos países estudam maneiras de se detectar este tipo de fraude, e muitas são as abordagens envolvendo Inteligência Artificial e Ciência de Dados. Como pode ser visto em CGU (2016), a Controladoria Geral da União (CGU) participa de congressos na área, inclusive em eventos internacionais, como a *Conference on Knowledge Discovery and Data Mining*, realizado em São Francisco, Califórnia, no ano de 2016.

Além disso, cabe destacar que a CGU possui uma unidade especializada em aplicar esse tipo de metodologia científica no combate à corrupção com dinheiro público, denominada Observatório de Despesas Públicas (ODP). O sucesso da iniciativa foi tanto, que a unidade já foi premiada diversas vezes em âmbito nacional, e ainda foi contemplada com o prêmio internacional *United Nations Public Service Awards*, em 2011 (CGU, 2019a).

Sendo assim, o presente artigo visa propor uma nova metodologia de combate à formação de cartéis, utilizando Mineração de Dados (MD) com Regras de Associação (RA). Esta nova abordagem foi inspirada em Silva (2011), e é obtida através de duas ações:

- Simplificação do método original, com eliminação da tarefa de *Clusterização* e os Sistemas de Multi-Agentes;
- Proposta de novo método para avaliar as muitas Regras de Associação (RA) geradas pelo algoritmo Apriori, com eliminação daquelas que não contribuem para indicação de suspeita de cartel.

3.2 REVISÃO BIBLIOGRÁFICA

Com a rápida evolução dos recursos computacionais vivida nos últimos anos, o volume de dados gerados e armazenados é cada vez maior e mais barato. Como pode ser visto em Gantz e Reinsel (2012), a estimativa para o crescimento da geração de dados é exponencial, e chegará a quatrocentos mil *exabytes* em 2020. Além disso, o mesmo estudo aponta que custo de armazenamento de dados digitais irá cair dez vezes entre os anos de 2012 e 2020.

Contudo, essa imensidão de dados digitais por si só, não é sinônimo de conhecimento, havendo assim a necessidade da criação de uma ferramenta capaz de processar essas informações, e extrair delas conhecimento útil. Em resposta a essa demanda, desponta nos anos noventa a Descoberta de Conhecimento em Base de Dados (DCBD), comumente chamada de Mineração de Dados (MD), que na realidade é apenas uma de suas etapas (SFERRA; CORRÊA, 2003).

Como exemplificado na figura 12, adaptada de Fayyad et al. (1996), a DCBD é constituída por cinco etapas, capazes de extrair padrões ocultos das bases de dados, com a finalidade de se obter conhecimento útil para a tomada de decisões. Esta tecnologia surgiu com a união entre três áreas afins: Estatística, Inteligência Artificial e Aprendizado de Máquinas, como afirmam Sferra e Corrêa (2003).

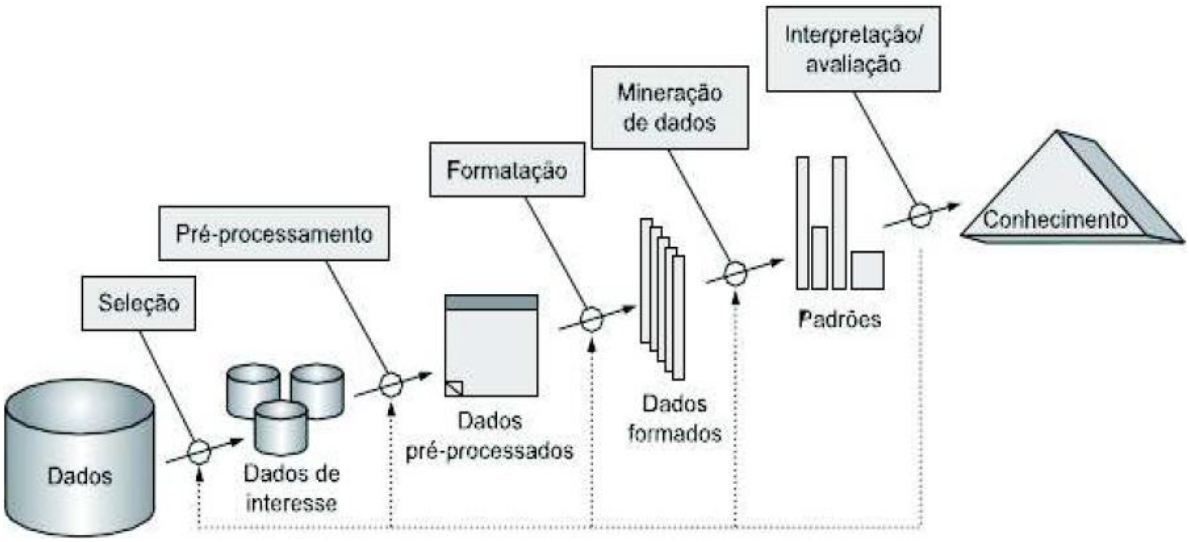

Figura 12 - Input, Etapas e Output da DCBD

Fonte: Adaptada de Fayyad *et al.* (1996).

Segundo Fayyad *et al.* (1996), a DCBD é dividida em: Seleção, Pré-Processamento, Formatação, Mineração de Dados (MD) e Interpretação dos resultados. Tais etapas estão definidas a seguir, com exemplos para o seu melhor entendimento.

- **Seleção:** Etapa em que são escolhidos os dados que irão ser utilizados durante o processo, eliminando assim aqueles que não são úteis, ou são redundantes. Por exemplo, caso haja o CNPJ e o nome fantasia das empresas em uma base, este último poderia ser excluído, deixando apenas o primeiro atributo (FAYYAD *et al.* 1996);
- **Pré-Processamento e Formatação:** Etapas com a finalidade de adequar a base de dados de modo que esta possa ser minerada, eliminando ruídos, itens incorretos, em branco, entre outros. Eliminar empresas com apenas uma participação em licitações públicas pode ser um exemplo de pré-processamento de dados, conforme o item 3.3.2 (FAYYAD *et al.* 1996);
- **Mineração de Dados (MD):** Etapa mais importante do processo, onde de fato os algoritmos serão executados na busca pelo conhecimento oculto. Existem diversas tarefas de MD com objetivos distintos, e técnicas diferentes para atingi-los. A Classificação, por exemplo, se utiliza da base de dados para treinamento de máquina, de modo que quando surge um novo registro, ela é capaz de predizer determinado atributo alvo, por meio de técnicas como a árvore de decisão. Já a *Clusterização* é uma tarefa que visa agrupar conjuntos de dados por sua similaridade, utilizando técnicas que medem a distância entre objetos em um espaço multiplano, constituído por eixos de todas as variáveis envolvidas. Como último exemplo, cabe citar a Associação, que visa identificar quando a ocorrência de determinada variável tem relação com outra. Como esta será a tarefa utilizada por esta metodologia, seus detalhes serão descritos mais adiante, juntamente com o respectivo algoritmo adotado (GOLDSCHMIDT; PASSOS, 2005);
- **Interpretação e Avaliação dos resultados:** Por fim, a última etapa avalia os resultados obtidos quanto a sua exatidão, qualidade, e verifica se estes são úteis à tomada de decisão. Cabe lembrar que há possíveis retornos às etapas anteriores, visando melhorias do produto final obtido (SUMATHI; SIVANANDAM, 2006);

Dentre as muitas formas de se minerar dados encontra-se a Associação, que segundo Camilo e Silva (2009), busca identificar a relação entre as variáveis, apresentando-se na forma de regras do tipo: “SE ocorre A, ENTÃO ocorre B”. Como pode ser visto em Larose (2005), em estudo realizado por uma rede de supermercados: SE o dia é sexta-feira, o período é noturno, e os clientes compram fraldas, ENTÃO eles compram cerveja. Pode-se perceber com esse exemplo como funcionam as Regras de Associação (RA), e a correlação criada entre as variáveis.

Como afirma Silva (2011), essas regras são criadas de acordo com valores mínimos de Suporte e Confiança, que são medidas de qualidade. Para explicar esses parâmetros, vamos escrever a regra apresentada no parágrafo anterior, e supor uma base de dados com apenas seis registros, conforme tabela 4.

Dia = Sexta; Período = Noite; Frauda = Sim → Cerveja = Sim.

Tabela 4 - Base de dados exemplo com seis registros

REGISTRO	DIA	PERÍODO	FRAUDA	(...)	CERVEJA
1	SEGUNDA	NOITE	NÃO	(...)	NÃO
2	QUARTA	TARDE	SIM	(...)	SIM
3	SEXTA	TARDE	NÃO	(...)	SIM
4	SEXTA	NOITE	SIM	(...)	SIM
5	TERÇA	MANHÃ	NÃO	(...)	NÃO
6	SEXTA	NOITE	SIM	(...)	NÃO

Fonte: Próprio Autor.

Observe que o lado esquerdo da regra aparece em dois registros de um total de seis, como ilustra a tabela 5. Sendo o Suporte a frequência relativa de aparição da “causa” da regra, podemos calculá-lo dividindo 2 por 6, e obtendo 33,33%.

Tabela 5 - Exemplo de cálculo do Suporte

REGISTRO	DIA	PERÍODO	FRAUDA	(...)	CERVEJA
1	SEGUNDA	NOITE	NÃO	(...)	NÃO
2	QUARTA	TARDE	SIM	(...)	SIM
3	SEXTA	TARDE	NÃO	(...)	SIM
4	SEXTA	NOITE	SIM	(...)	SIM
5	TERÇA	MANHÃ	NÃO	(...)	NÃO
6	SEXTA	NOITE	SIM	(...)	NÃO

Fonte: Próprio Autor.

Como pode ser visto na tabela 6, dentre as duas vezes que ocorreu a causa, em apenas uma delas a consequência foi verdadeira, ou seja, houve compra de cerveja. Sendo a Confiança a frequência relativa de ocorrência do lado direito da regra, dado que já ocorreu o lado esquerdo, podemos calculá-la dividindo 1 por 2, e obtendo 50%.

Tabela 6 - Exemplo de cálculo da Confiança

REGISTRO	DIA	PERÍODO	FRAUDA	(...)	CERVEJA
1	SEGUNDA	NOITE	NÃO	(...)	NÃO
2	QUARTA	TARDE	SIM	(...)	SIM
3	SEXTA	TARDE	NÃO	(...)	SIM
4	SEXTA	NOITE	SIM	(...)	SIM
5	TERÇA	MANHÃ	NÃO	(...)	NÃO
6	SEXTA	NOITE	SIM	(...)	NÃO

Fonte: Próprio Autor.

Tendo entendido os conceitos de Suporte e Confiança, pode-se entender o funcionamento dos algoritmos de geração dessas regras. Como explicado em Tan, Steinbach e Kumar (2009), é inviável a geração de todas as Regras de Associação (RA) possíveis e cálculo dos seus respectivos Suportes e Confianças, quando se trata de bases de dados mais volumosas. Isso ocorre porque a complexidade do problema é de $3^d - 2^d + 1$, onde “d” é o número de itens da base de dados. Como solução a este problema, foi proposto por Agrawal e Srikant (1994) o algoritmo Apriori, que é o mais utilizado e influente nessa área, e possui estratégias inteligentes de poda por Suporte e Confiança, reduzindo consideravelmente o tempo de processamento computacional.

APRIORI: PODA BASEADA NO SUPORTE:

Como explicam Agrawal e Srikant (1994), se um conjunto é frequente na base de dados, então todos os seus subconjuntos também serão. Em outras palavras, se {a, b, c} aparece frequentemente nos registros, então seus subconjuntos {a, b}, {a, c}, {b, c}, {a}, {b} e {c} também serão frequentes. Em contrapartida, se um conjunto {a, b} não é frequente, seus superconjuntos {a, b,...} também não serão.

Entendido este princípio, a partir de um conjunto que não atenda ao requisito de Suporte Mínimo, pode-se eliminar todos os seus superconjuntos sem necessidade de cálculo de Suporte. Observe que com este princípio o algoritmo Apriori faz uma poda inteligente, que reduz consideravelmente a complexidade do problema, permitindo assim sua aplicação em grandes bases de dados.

APRIORI: PODA BASEADA NA CONFIANÇA:

Na realização da poda por valores de Confiança Mínima, o algoritmo Apriori gera primeiramente regras cujo consequente contenha apenas um item, por exemplo: {a, c, d} → {b} e {a, b, d} → {c}. Caso elas apresentem Confiança satisfatória, acontece a fusão das mesmas, gerando uma nova regra: {a, d} → {b, c}. A nova Confiança é calculada, e se ela atende ao mínimo, ela é mantida.

Em contrapartida, se uma regra com único consequente possui Confiança baixa, então todas as regras com o mesmo consequente são eliminadas, reduzindo consideravelmente a complexidade do problema. Ou seja, se {b, c, d} → {a} não atende a Confiança Mínima, logo {c, d} → {a, b} e {b, d} → {a, c} também não, e são excluídas.

3.3 METODOLOGIA

Uma característica marcante dos cartéis é o fato de que as empresas concorrem sempre juntas nos processos de licitação. Sendo assim, se a empresa “A” está participando de uma licitação, e ela pertence a determinado cartel, as chances das demais empresas pertencentes a este conluio também estarem inscritas neste mesmo processo licitatório é alta (CAMILO; SILVA, 2009).

Sendo assim, inspirado em Silva (2011), este trabalho propõe uma metodologia que utiliza a Descoberta de Conhecimento em Base de Dados (DCBD), onde a etapa de Mineração de Dados (MD) é feita por meio das chamadas Regras de Associação (RA). Além disso, é proposto neste artigo um novo algoritmo de ranqueamento do grande volume de regras obtidas, facilitando assim a interpretação dos resultados. Os tópicos a seguir detalham a metodologia proposta, alinhada com as etapas do DCBD mencionadas no referencial teórico.

3.3.1 SELEÇÃO DOS DADOS

Nesta etapa deve-se obter uma base de dados com o CNPJ dos participantes das licitações de determinado objeto específico, como exemplo: serviços de engenharia, peças automotivas, materiais de construção, entre outros. Feito isso, os demais atributos devem ser descartados, dentre eles: as informações sobre o órgão

licitante, a razão social ou nome fantasia da empresa, ou o número da licitação (SILVA, 2011).

3.3.2 PRÉ-PROCESSAMENTO E FORMATAÇÃO

Para execução futura da Mineração de Dados (MD), é necessário que a base seja transformada conforme indicado da tabela 7. Observe que as linhas representam as licitações, e as colunas os possíveis fornecedores. Neste caso hipotético, a empresa 1 participa da licitação 6, enquanto a empresa 2 participa das licitações 2 e 4 (SILVA, 2011).

Tabela 7 - Exemplo da Base de Dados

	Empresa 1	Empresa 2	Empresa 3	Empresa 4	Empresa 5
Licitação 1					
Licitação 2		SIM			
Licitação 3			SIM		
Licitação 4		SIM			
Licitação 5					
Licitação 6	SIM			SIM	

Fonte: Próprio autor.

Feito isso, de acordo com Silva (2011), algumas outras ações de Pré-Processamento e Formatação são necessárias para que o algoritmo de MD funcione corretamente, e traga bons resultados. São eles:

- Eliminação de linhas repetidas, que podem surgir com o duplo cadastro de determinada licitação na base de dados;
- Eliminação de fornecedores que tenham participado de apenas uma licitação ao longo de toda base. Isso ocorre porque esta metodologia foi criada justamente para detectar empresas que tendem a entrar juntas em licitações, logo, não faz sentido analisar aquelas que participaram de editais em apenas em uma ocasião.
- Eliminar licitações com apenas um fornecedor, que podem surgir nos casos de licitação inexigível ou dispensada. Essa exclusão é justificada pelo fato

de que os cartéis são fraudes realizadas com a participação de grupos de empresas, logo, a participação de um único fornecedor não será analisada.

3.3.3 MINERAÇÃO DE DADOS

Conforme mencionado no referencial teórico, nesta etapa são efetivamente analisados os dados com algoritmos que visam descobrir padrões ocultos. Já que o intuito é descobrir quando as empresas tendem a participar das licitações em conjunto, a tarefa de Associação é a mais indicada dentre as muitas opções de Mineração de Dados (SFERRA; CORRÊA, 2003).

A ideia do estudo é descobrir regras, como a apresentada a seguir: “A = Sim”, “B = Sim” → “C = Sim” (Confiança = 80%). Note que, caso as empresas “A” e “B” estejam participando de uma licitação, as chances da empresa “C” também estar são de 80%, o que evidencia indício de cartel entre elas.

Neste ponto cabe uma discussão interessante sobre o parâmetro de “poda” chamado Suporte, utilizado na geração dessas regras. Observe que, como afirma Silva (2011), valores altos de suporte não garantem boas regras, e inclusive contribuem para a omissão de outras reveladoras. Isso ocorre porque o alto suporte indica que certo grupo de empresas participou de muitas licitações em conjunto, o que é muito comum quando se trata de fornecedores de grande porte, e não necessariamente quer dizer uma fraude.

Em contrapartida, pode haver grupos de pequenas empresas que participam de apenas 15 licitações em toda base de dados, entretanto quase sempre esses editais foram concorridos em conjunto com outras empresas, caracterizando um indício de fraude. Em vista disso, é proposto neste trabalho que seja adotado um valor de Suporte Mínimo baixo, por exemplo, de 0,1%, evitando a perda de regras importantes (SILVA, 2011).

Cabe destacar, conforme Silva (2011), que o baixo valor de Suporte Mínimo adotado acarreta a geração de muitas Regras de Associação (RA). Em vista disso, é necessário um algoritmo auxiliar capaz de ranqueá-las por outro critério mais adequado, como descrito no tópico 3.3.4.

Por fim, relativo à Confiança Mínima, foi utilizado o valor de 90%, como proposto em Silva (2011). Este percentual foi discutido por especialistas do assunto, e definido como um valor plausível para suspeita da formação de um cartel.

3.3.4 INTERPRETAÇÃO E AVALIAÇÃO

Como mencionado no item 3.3.3, com o baixo Suporte Mínimo adotado, é inevitável um alto número de Regras de Associação (RA). Sendo assim, é necessária a criação de algum critério capaz de orientar os esforços humanos para averiguação apenas das regras mais relevantes. Neste ponto, este artigo difere da metodologia de Silva (2011), e propõe um novo Algoritmo de Seleção Regras.

Para explicar esta proposta, far-se-á uso do seguinte exemplo: “A = Sim”, “B = Sim”, “C = Sim” → “D = Sim” (Confiança = 93%). É importante notar que as empresas “A”, “B” e “C” estão presentes na “causa” da regra, enquanto a empresa “D” representa a “consequência”. Outro fato a ser percebido é a frequência de acerto da regra, representado por uma Confiança de 93%.

Caso o número de licitações envolvendo “A”, “B” e “C” fosse 15, logo, o número de licitações envolvendo “A”, “B”, “C” e “D” seria de 14, originando assim uma Confiança de 93% (14/15). De forma análoga, esta Regra de Associação (RA) poderia ser representada pelo conjunto e subconjunto apresentado na figura 13.

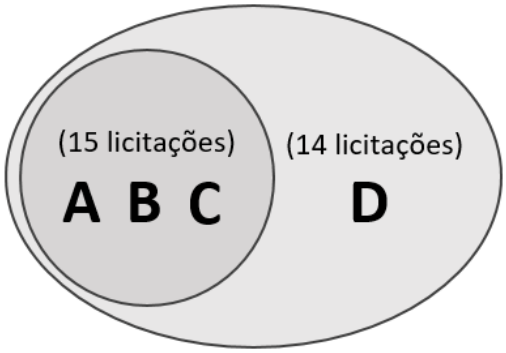

Figura 13 - Regra de Associação vista como conjunto e subconjunto.
Fonte: Próprio autor.

Este trabalho propõe que o Algoritmo de Seleção de Regras classifique-as de acordo com quatro situações distintas. Nas situações 1 e 2, as regras são descartadas da análise, pois não indicam a formação de cartel. Na situação 3, as regras são aproveitadas parcialmente, eliminando a parte que contem empresas livre de suspeita. Por fim, a situação 4 implica em considerar a regra por inteiro, com indício de todas as empresas estarem praticando a fraude.

SITUAÇÃO 1: Situação em que todas as empresas têm Participação Relativa (PR) baixa, em virtude do pequeno número de participações conjuntas, se comparado às participações individuais. Como exemplo, pode-se imaginar que a empresa “A” tenha participado de 140 licitações, e apenas 14 delas foi em conjunto com as demais, logo, sua Participação Relativa (PR) é de 10% (14/140). A figura 14 ilustra o que foi dito, considerando participações relativas de 10%, 10%, 15% e 25% para as empresas “A”, “B”, “C” e “D”, respectivamente.

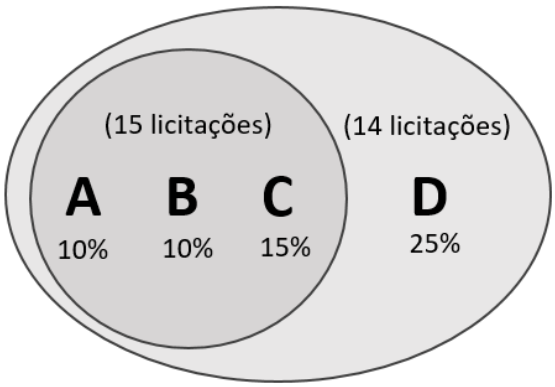

Figura 14 - Primeira situação
Fonte: Próprio autor.

É importante observar que, apesar da regra suposta ter um alto nível de Confiança (93%), ela não pode ser considerada um indício de fraude. Isso ocorre porque todas essas empresas participam de muitas licitações, tornando o número de participações delas em conjunto (14 licitações) relativamente baixo. Em vista disso, este trabalho propõe a exclusão das regras cuja Participação Relativa (PR) é baixa para todas as empresas integrantes.

SITUAÇÃO 2: Trata-se do mesmo caso anterior, porém com a participação de “A” em 15 licitações. Isso quer dizer que 100% das vezes que “A” participou, ela estava em conjunto com as demais, conforme representado na figura 15. Nessa situação, se entende que “B”, “C” e “D” são três grandes empresas, que figuram essa regra apenas por concorrerem a muitas licitações, enquanto “A” é uma pequena empresa que concorreu sempre à sombra das rivais de grande porte. Em vista disso, este trabalho também propõe que, quando apenas uma das empresas tem alto valor de Participação Relativa (PR), a regra seja descartada.

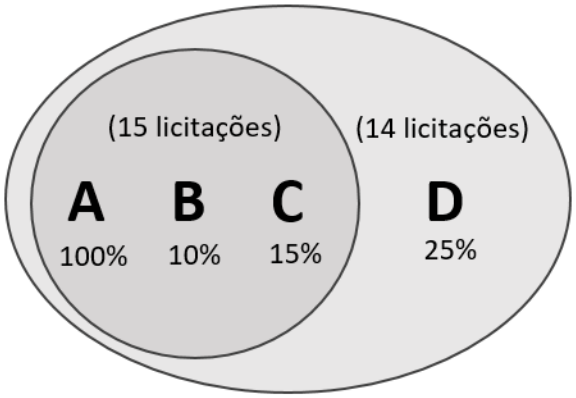

Figura 15 - Segunda situação.
Fonte: Próprio autor.

SITUAÇÃO 3: Neste caso, duas ou mais empresas têm alto índice de Participação Relativa (PR), enquanto as demais apresentam baixa participação, como representado na figura 16. É importante observar que “B” e “C” são empresas que concorreram a muitas outras licitações, sendo as 14 em conjunto, irrelevantes para suspeita. Já as empresas “A” e “D” tem no número 14, um alto percentual de suas licitações, o que indica que estas duas empresas concorrem juntas na maioria das vezes, sugerindo assim um cartel. Em vista disso, nessa situação, este trabalho propõe eliminar as empresas com baixa Participação Relativa (PR) da regra (B e C), e manter a suspeita sobre as demais (A e D).

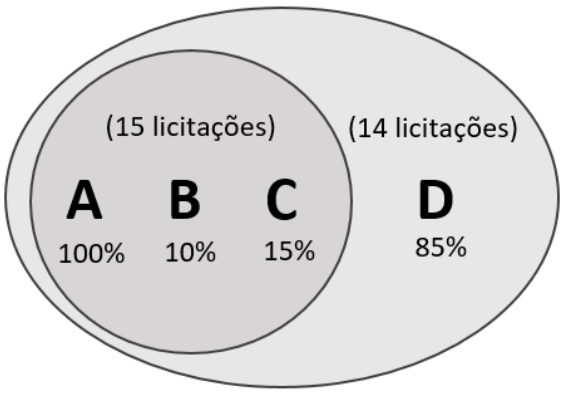

Figura 16 - Terceira situação
Fonte: Próprio autor.

SITUAÇÃO 4: Este último caso ocorre quando todas as empresas pertencentes à determinada regra possuem alto índice de Participação Relativa (PR), como ilustra a

figura 17. Nesta situação, as 14 licitações que elas participam em conjunto são representativas se comparadas à atuação de cada uma delas individualmente. Isso indica que, quando elas concorrem, quase sempre estão juntas, o que sugere a formação de um cartel. Nesse caso, a regra é aproveitada por inteiro, estando “A”, “B”, “C” e “D”, suspeitas de formar um conluio.

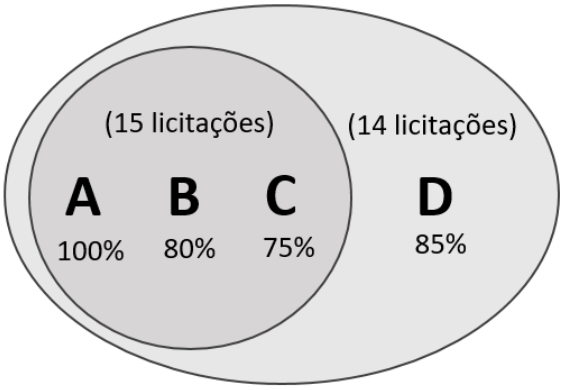

Figura 17 - Quarta situação.
Fonte: Próprio autor.

Feito isso, cabe ao executor do método analisar as regras remanescentes, após aplicação do algoritmo de poda proposto, e avaliar se há ou não a formação de cartéis. Fato bem interessante a ser comentado, é o valor limite para se considerar que a Participação Relativa (PR) é alta ou baixa dentro do funcionamento do algoritmo. Este valor não foi fixado por esse trabalho, pois se acredita que este deve ser avaliado caso a caso, por auditores especialistas, e testado em estudo de caso, com uma base de dados real, com um objeto de licitação definido, a fim de se chegar a um valor razoável.

Cabe também dizer que o número de regras remanescentes após execução do algoritmo aumenta quando este limite é diminuído. A título de exemplo, na situação 3 a regra foi considerada válida, porque 14 licitações correspondiam a 85% dos editais da empresa “D”. Caso esses 85% não fossem considerados como suficientes para suspeitar, essa regra teria sido eliminada pelo algoritmo.

3.4 CONCLUSÃO

O objetivo deste artigo foi de propor uma nova metodologia de análise de carteis em licitações públicas. Após Introdução e breve Revisão Bibliográfica, a Metodologia

proposta foi descrita em detalhes, desde a Seleção dos dados, passando pelo Pré-Processamento e Formatação dos mesmos, seguido da Mineração de Dados (MD) por meio de Regras de Associação (RA) com algoritmo Apriori, e concluindo com a descrição de um novo Algoritmo de Seleção de Regras.

Vale frisar que este algoritmo foi proposto no presente artigo, e detalhado em quatro situações, com exemplos práticos envolvendo empresas “A”, “B”, “C” e “D” para melhor entendimento de seu funcionamento. Cabe a trabalhos futuros dar continuidade a esse estudo, aplicando o método a uma base de dados real de licitações, a fim de verificar sua eficiência e eficácia, e descobrir possíveis deficiências e pontos de melhorias.

3.5 REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. *In:* INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994. **Proceedings** […]. Hove, East Sussex: Morgan Kaufmann, 1994. p. 487-499.

CAMILO, C.; SILVA, J. **Mineração de Dados:** Conceitos, Tarefas, Métodos e Ferramentas. Goiânia: Universidade Federal de Goiás, 2009.

CGU. **Observatório da Despesa Pública participa de congresso internacional sobre mineração de dados**. 2016. Disponível em: https://www.cgu.gov.br/noticias/2016/08/observatorio-da-despesa-publica-participa-de-congresso-internacional-sobre mineracao-de-dados. Acesso em: 5 ago. 2019.

CGU. **Observatório da Despesa Pública**. 2019a. Disponível em: https://www.cgu.gov.br/assuntos/informacoes-estrategicas/observatorio-da-despesa - publica. Acesso em: 5 ago. 2019.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Advances in knowledge discovery and data mining**. Menlo Park, Calif.: AAAI Press, USA, 1996.

GANTZ, J.; REINSEL, D. **The Digital Universe In 2020:** Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. IDC Corporate, Framingham, dec. 2012. Disponível em: https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf. Acesso em: 14 jun. 2019.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. São Paulo: Elsevier, 2005.

LAROSE, D. T. **Discovering knowledge in data:** an introduction to data mining. Hoboken, N. J.: Wiley-Interscience, 2005.

PIETRO, M. S. Z. D. **Direito Administrativo**. 32. ed. São Paulo: Forense, 2019.

SANTOS, F. B.; SOUZA, K. R. de. **Como combater a corrupção em Licitações**. 2. ed. Belo Horizonte: Fórum, 2018.

SFERRA, H. H.; CORRÊA, A. C. J. Conceitos e Aplicações de Data Mining. **Revista de Ciência & Tecnologia,** Piracicaba, v. 11, n. 22, p. 19–34, 2003.

SILVA, C. V. S. **Agentes de Mineração e sua Aplicação no Domínio da Auditoria Governamental**. Brasília: Universidade de Brasília, 2011.

SUMATHI, S.; SIVANANDAM, S. N. **Introduction to data mining and its applications**. Berlin; New York: Springer, 2006.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

4 ARTIGO C - MINERAÇÃO DE DADOS NO COMBATE AOS CARTÉIS: ESTUDO DE CASO

Resumo

Um dos principais problemas enfrentados por governos do mundo todo no que tange a compras de bens ou contratação de serviços são os chamados cartéis em licitações públicas. Trata-se de um acordo prévio realizado por um grupo de empresas, que se revezam na posição vencedora, eliminando assim a competição e aumentando os preços para os cofres públicos. O objetivo deste artigo é aplicar Mineração de Dados (MD) no combate à formação de cartéis, por meio da Tarefa de Associação, metodologia proposta por Silva (2011), aplicada com alterações apresentadas neste artigo. O estudo de caso foi realizado com uma base de dados aberta, disponibilizada no sítio eletrônico da Controladoria Geral da União (CGU), adotando um objeto específico dentre os bens e serviços adquiridos pela administração pública. Como não é objetivo deste trabalho oferecer uma denúncia à CGU, e sim testar a metodologia proposta em caráter experimental, o CNPJ das empresas participantes foi substituído por números aleatórios de quatro dígitos, preservando assim em sigilo, a identidade das empresas envolvidas.

Palavras-chave: Mineração de Dados. DCBD. Licitações. Cartéis.

Abstract

One of the main problems faced by governments around the world, related to purchases or hiring of services are the cartels in the public biddings. This is a prior agreement involving a group of companies, rotating the winning position, eliminating the competition and increasing the prices to the public administration. The purpose of this article is to apply Data Mining (DM) in the fighting against the cartels, using Association Rules, methodology proposed by Silva (2011), applied with changes described in this article. The study of case was conducted with an open database, available on the website of *Controladoria Geral da União* (CGU), adopting a specific object among the goods and services purchased by the public administration. As the purpose of this paper is not to submit a complaint to the CGU, but to test the proposed

methodology on an experimental basis, the real ID of the participating companies was replaced by random four-digit numbers, preserving the identity of the companies involved in confidentiality.

Keywords: Data Mining. KDD. Bidding. Cartels.

4.1 INTRODUÇÃO

Para que o governo realize suas compras com Legalidade, Impessoalidade, Moralidade, Publicidade e Eficiência, este realiza um processo administrativo denominado de licitação, conforme ressalta Pietro (2019). Trata-se de um conjunto de procedimentos definidos em lei, de modo a garantir que os fornecedores possam competir de maneira justa entre si, garantindo ao mesmo tempo, o não favorecimento de nenhum particular, e também os melhores preços aos cofres públicos (NIEBUHR; NIEBUHR, 2018).

Contudo, há uma fraude que assombra os governos de todo o mundo no que diz respeito a compras públicas, denominada cartel. Trata-se de um acordo prévio entre os fornecedores participantes do edital, que se revezam na posição vencedora fazendo um rodízio, eliminando a competição, e aumentando assim os preços pagos pelo governo (SANTOS; SOUZA, 2018).

Por essa motivação, esse artigo visa aplicar a Descoberta de Conhecimento em Base de Dados (DCBD) a uma base de dados aberta de licitações, disponibilizada pela Controladoria Geral da União (CGU) em seu sítio eletrônico, visando testar uma metodologia capaz de revelar a formação de cartéis. Essa proposta foi inspirada em Silva (2011), com algumas alterações que serão apresentadas neste artigo.

4.2 REVISÃO BIBLIOGRÁFICA

Como explicam Sferra e Corrêa (2003), com o grande volume de dados digitais gerados pelos sistemas informatizados, o homem percebeu a oportunidade de extrair conhecimento útil dos mesmos, e a ferramenta capaz de fazê-lo ficou conhecida como Descoberta de Conhecimento em Base de Dados (DCBD). De acordo com Fayyad et al. (1996), trata-se de um processo não trivial, estruturado em cinco etapas, interativo

e iterativo, que busca extrair padrões ocultos nos bancos de dados, que sejam novos e potencialmente úteis para a tomada de decisões.

Como pode ser observada na figura 18, adaptada de Fayyad et al. (1996), as cinco etapas da DCBD são: Seleção, Pré-Processamento, Formatação, Mineração de Dados (MD) e Interpretação de Resultados. Conforme explicam Sumathi e Sivanandam (2006), elas podem ser traduzidas como:

- **Seleção:** Etapa em que os dados úteis ao estudo são selecionados, enquanto aqueles que não possuem importância são descartados;
- **Pré-processamento e Formatação:** Etapas de preparo da base de dados para a etapa seguinte. Compreende a adequação de itens incorretos, em branco, remoção de ruídos, formatação a padrões específicos de cada algoritmo de mineração, entre outros;
- **Mineração de Dados (MD):** Etapa mais importante do processo, onde algoritmos são executados na busca por padrões ocultos na base de dados. Dentre as muitas Tarefas da mineração, existe a Classificação, a *Clusterização*, a Associação, entre outras;
- **Interpretação e Avaliação dos Resultados:** Etapa conclusiva, onde se avalia se os padrões obtidos são úteis ou não para tomada de decisões. Cabe destacar que a análise deste produto final pode sugerir no retorno a alguma etapa anterior, a fim de se obter melhorias.

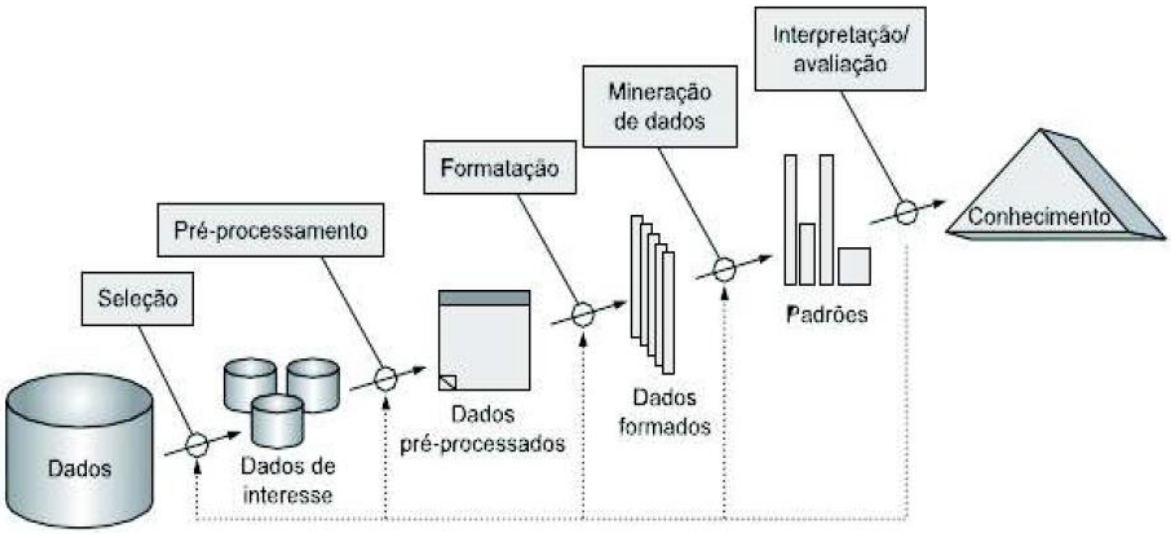

Figura 18 - Input, Etapas e Output da DCBD

Fonte: Adaptada de Fayyad et al. (1996).

Como mencionado, uma das mais importantes tarefas de Mineração de Dados (MD) é a chamada Associação, que visa identificar a relação entre variáveis com regras de causa e consequência (SE... ENTÃO...), conforme explicam Camilo e Silva (2009).

Exemplificando com este estudo de caso, poderíamos encontrar regras do tipo: SE a empresa “A” e a empresa “B” participam de uma licitação, ENTÃO a empresa “C” também participa. Observe que esta é uma situação em que as três empresas tendem a concorrer juntas, logo, há certa suspeita de formação de cartel. Reescrevendo a regra no padrão formal, ficaria: “A = Sim”, “B = Sim” → “C = Sim”.

Como explica Larose (2005), essas regras possuem medidas de qualidade, chamadas de Suporte e Confiança. A primeira diz respeito ao percentual de vezes que a “causa” (lado esquerdo da regra) aparece na base de dados, já a segunda indica o percentual de vezes que a “consequência” (lado direito da regra) é verdadeira, caso ocorra a “causa”. Para ilustrar, basta supor que a base de dados possui 10.000 registros, e que as empresas “A” e “B” apareçam juntas por 20 vezes, logo, o Suporte dessa regra é obtido dividindo 20 por 10.000 (0,2%). Se, dentre essas 20 aparições conjuntas de “A” e “B”, a empresa “C” aparecesse 15 vezes, então a Confiança seria obtida dividindo 15 por 20 (75%).

Entendidos esses conceitos iniciais, pode-se entender as ferramentas usadas na Associação. De acordo com Goldschmidt e Passos (2005), os algoritmos que produzem as chamadas Regras de Associação (RA) recebem como *input* valores mínimos de Suporte e Confiança, a fim de gerar regras que se enquadrem nesses dois requisitos. Como explica Tan, Steinbach e Kumar (2009), quando a base de dados é muito volumosa, é inviável calcular esses dois parâmetros para todas as regras possíveis, já que a complexidade é de $3^d - 2^d + 1$, onde “d” é o número de itens da base de dados.

Nesse contexto, em 1994 foi proposto por Agrawal e Srikant (1994) o algoritmo Apriori. Esta ferramenta usa de estratégias que evitam o cálculo exaustivo de todas as regras, reduzindo assim consideravelmente o tempo de processamento. O funcionamento do algoritmo em mais detalhes pode ser encontrado em Tan, Steinbach e Kumar (2009), e será brevemente explicado a seguir.

PODA BASEADA NO SUPORTE:

O Primeiro artifício usado pelo algoritmo Apriori que reduz consideravelmente o tempo de processamento das regras se baseia na seguinte premissa: Se um conjunto é frequente, então todos os seus subconjuntos também serão, e analogamente, se um subconjunto não é frequente, então seus superconjuntos também não serão.

A partir dessa afirmação, se um subconjunto {a, b} não é frequente, o algoritmo elimina todas as regras contendo seus superconjuntos {a, b,...}, sem necessidade de mais cálculos. É importante perceber que essa estratégia é capaz de reduzir consideravelmente o volume de cálculos executados pelo computador.

PODA BASEADA NA CONFIANÇA:

Além disso, o algoritmo possui uma estratégia inteligente para realizar a poda pelo valor mínimo de Confiança. Primeiramente o Apriori se utiliza de regras com apenas um elemento consequente, como exemplo: {a, c, d} → {b} e {a, b, d} → {c}. Caso elas apresentem Confiança satisfatória, ocorre a fusão das mesmas, gerando uma nova regra: {a, d} → {b, c}. A nova Confiança é calculada, e se atendido o valor mínimo, ela é mantida.

De maneira análoga, se a referida regra possui Confiança não satisfatória, então todas as regras contendo aquele consequente também são descartadas, reduzindo consideravelmente a complexidade do problema. Em outras palavras, se {b, c, d} → {a} tem baixa Confiança ela será eliminada, e consequentemente, as regras {c, d} → {a, b} e {b, d} → {a, c} também serão.

4.3 CLASSIFICAÇÃO DA PESQUISA

A pesquisa será classificada segundo seus objetivos de acordo com Jung (2003), como Básica ou Aplicada. A primeira adquire conhecimento, porém sem aplicação prática e imediata. Já a segunda, utiliza ferramentas e a literatura disponível para produzir aplicações práticas. Como este trabalho visa testar uma metodologia para aplicações futuras, esta pesquisa se classifica como Básica.

Nesta mesma linha, a pesquisa também pode ser classificada como: Descritiva, Explicativa ou Exploratória. A primeira se limita em registrar e analisar o fenômeno, sem entrar no mérito de seu conteúdo. Já a segunda, como o nome já diz, visa explicar a ocorrência daqueles fatos, trazendo a luz um conhecimento de causa e

consequência. Por fim, a Exploratória, como se classifica essa pesquisa, é a que busca investigar, formular questões e propor hipóteses, abrindo portas para pesquisas futuras mais sólidas e consistentes (LAKATOS; MARCONI, 2017).

Outra classificação que pode ser feita é referente à abordagem utilizada: Qualitativa, Quantitativa, ou ambas. A primeira não se preocupa em representar os resultados de forma numérica, e a compreensão é feita através de classificações, grupos, entre outros. Já a segunda, traduz as informações coletadas em números, que podem ser analisados ou classificados. Como se trata de Mineração de Dados (MD), a característica Quantitativa é inerente ao funcionamento dos algoritmos, e também refletida nos resultados obtidos pelos parâmetros Suporte e Confiança. Entretanto, também há elementos Qualitativos, principalmente na interpretação de resultados por parte do auditor especialista, o que sugere que esta seja uma pesquisa que envolva ambas as abordagens (GERHARDT; SILVEIRA, 2009).

Por fim, a última classificação apresentada será com relação ao procedimento técnico usado, que pode ser: Pesquisa Bibliográfica, Pesquisa de Campo ou Estudo de Caso. A primeira consiste em fazer um apanhado na bibliografia já existente, reunindo as publicações relevantes acerca do tema. Já a segunda se relaciona com a observação e coleta de dados julgados relevantes, a fim de se estudar futuramente determinado fenômeno, com base na análise dos mesmos. Por fim, a última é a investigação empírica de um fenômeno contemporâneo que ocorre no mundo real. Observe que, como está sendo estudada a formação de cartéis nas licitações públicas, esta pesquisa se encaixa como um Estudo de Caso (YIN, 2014).

4.4 METODOLOGIA COM ESTUDO DE CASO

4.4.1 BASE DE DADOS

A base de dados utilizada está disponível no portal transparência da CGU, visto em CGU (2019b), acessado em 01 de julho de 2019, com licitações abrangendo um período de cinco anos. Na tabela 8, adaptada da base original com a omissão de certas informações, podem ser observados os atributos: números dos processos, informações sobre órgãos licitantes, unidades de destino, bens ou serviços adquiridos, fornecedores participantes e vencedores.

Tabela 8 - Exemplo adaptado da base de dados utilizada.

Núm. Proc.	Cód. Órgão	Nome Órgão	Cód. UG	Nome UG	Cód. Item	Descrição Item	CNPJ	Nome Particip.	Vencedor
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
A	B	C	D	E	1103...	MAN/REF PREDIAL	...	...	NÃO
F	G	H	I	J	9266...	MAT P/ VNI...	...	...	NÃO
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	NÃO
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	SIM
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	NÃO
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	NÃO
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	NÃO
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	NÃO
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	NÃO
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	SIM
F	G	H	I	J	9266...	MAR P/ VNI...	...	...	NÃO

Fonte: Próprio autor.

4.4.2 SELEÇÃO DOS DADOS

Com o intuito de preparar a base de dados para aplicação do algoritmo Apriori, foi necessário excluir certos atributos, conforme metodologia descrita em Silva (2011):

1) **Número do processo:** Este atributo é omitido para execução do algoritmo de mineração, porém permanece na base para fins de identificação da licitação.
2) **Nome e código do órgão licitante e do órgão de destino:** Nada impede que os cartéis participem em licitações nos mais variados órgãos, e por essa razão, essa informação não é relevante, e foi excluída.
3) **Nome e código do item comprado:** Conforme descrito no resumo, este estudo de caso foi feito adotando um objeto específico de compra. Sendo assim, todas as licitações envolvendo bens ou serviços distintos a este, foram excluídas.
4) **Nome fantasia e CNPJ do participante:** Devido ao caráter experimental da pesquisa, que visa testar a metodologia e não oferecer uma denúncia propriamente dita aos órgãos de controle, o nome fantasia e o CNPJ das empresas foram excluídos, e substituídos por um número de identificação aleatório de quatro dígitos.
5) **Identificação do Vencedor:** Este atributo não participa das Regras de Associação (RA), porém após a mineração ele será útil para verificar se os cartéis formados realmente obtiveram sucesso.

4.4.3 PRÉ-PROCESSAMENTO E FORMATAÇÃO

Conforme descrito em Silva (2011), foi gerado um algoritmo capaz de transformar a base de dados apresentada nos itens 4.4.1 e 4.4.2 em uma matriz, onde cada linha ou instância representa uma licitação, e cada coluna ou atributo representa uma empresa. Dessa forma, quando uma empresa participa de determinada licitação, a interseção entre essas linhas e colunas recebe a variável *booleana* “SIM”, e caso contrário, “NÃO”. Conforme exemplo da tabela 9, a empresa “A” participa da licitação “5555”, a “B” participa das licitações “2222” e “3333”, e assim por diante.

Tabela 9 - Exemplo de matriz com variáveis "SIM" e "NÃO"

		EMPRESAS				
		A	B	C	D	(...)
LICITAÇÕES	1111	NÃO	NÃO	NÃO	NÃO	...
	2222	NÃO	**SIM**	NÃO	**SIM**	...
	3333	NÃO	**SIM**	NÃO	**SIM**	...
	4444	NÃO	NÃO	NÃO	NÃO	...
	5555	**SIM**	NÃO	**SIM**	NÃO	...
	(...)	...	...	...	...	...

Fonte: Próprio autor.

Entretanto, percebeu-se que essa matriz não era satisfatória, pelo fato de o algoritmo de Associação gerar, em sua grande maioria, regras com a variável “NÃO”. Estas regras não trazem nenhuma evidência da formação de cartéis, logo, para solucionar este problema, o “NÃO” foi substituído pelo caractere “?”. Isso foi feito porque, na linguagem do *software* utilizado, isso significa uma variável não informada ou *missing*, e consequentemente, são geradas apenas as regras com a variável “SIM”. Na tabela 10, segue o exemplo da nova transformação.

Tabela 10 - Exemplo de matriz com variáveis "SIM" e “?"

		EMPRESAS				
		A	B	C	D	(...)
LICITAÇÕES	1111	?	?	?	?	...
	2222	?	**SIM**	?	**SIM**	...
	3333	?	**SIM**	?	**SIM**	...
	4444	?	?	?	?	...
	5555	**SIM**	?	**SIM**	?	...
	(...)	...	...	...	...	...

Fonte: Próprio autor.

Feito isso, ainda há três alterações que o algoritmo deve fazer de modo a preparar a base para a Mineração de Dados (MD), conforme listado a seguir:

- **Exclusão das linhas repetidas:** Foi constatado que havia uma série de linhas repetidas, em virtude de um possível cadastro duplo na base de dados, e estas foram removidas;

- **Exclusão de empresas com apenas uma licitação:** Quando uma empresa participa de apenas um processo licitatório, não há que se falar em formação de cartel. Sendo assim, ela é excluída;
- **Exclusão de licitações dispensadas ou inexigíveis:** Havia também na base, registros de licitações com apenas um participante. Neste caso ocorreu uma licitação inexigível ou dispensada, e logo, também não há que se falar em formação de cartel. Sendo assim, essas licitações foram removidas;

4.4.4 MINERAÇÃO DE DADOS

Para realização da Mineração de Dados (MD), etapa em que definitivamente os algoritmos são executados na busca por padrões ocultos, foi utilizado o pacote de *software* WEKA (*Waikato Environment for Knowledge Analysis).* Este *software* é de uso gratuito, possui código aberto, e foi desenvolvido inicialmente pela Universidade de *Waikato*, na Nova Zelândia. A escolha por este aplicativo se deu pelo fato de que esta ferramenta, além de possuir uma interface de fácil compreensão e uso, também contempla às necessidades do trabalho, abrangendo diversos algoritmos com a tarefa de Associação, dentre eles o Apriori (WAIKATO, 2019).

Como mencionado no referencial teórico, para que sejam geradas as Regras de Associação (RA), é necessário o fornecimento de dois parâmetros de entrada: o Suporte Mínimo e a Confiança Mínima. Conforme explicado por Silva (2011), altos valores de Suporte Mínimo irão excluir regras importantes, sendo assim, foi adotado 0,1% para esse parâmetro. Em consequência do tamanho da base de dados, este percentual corresponde a 14 aparições do conjunto de empresas que formam a "causa" da regra.

Com relação à Confiança Mínima, conforme definido em Silva (2011), foi adotado o valor de 90%, considerado por especialistas um percentual razoável. Este índice, pela dimensão da base de dados usada, equivale a um mínimo de 12 licitações conjuntas das empresas da figuram a "consequência" da Associação. A figura 19 e a regra a seguir ilustram o que foi dito nos dois últimos parágrafos.

"A = Sim"; "B = Sim" → "C = Sim". (Suporte = 0,1% e Confiança = 90%)

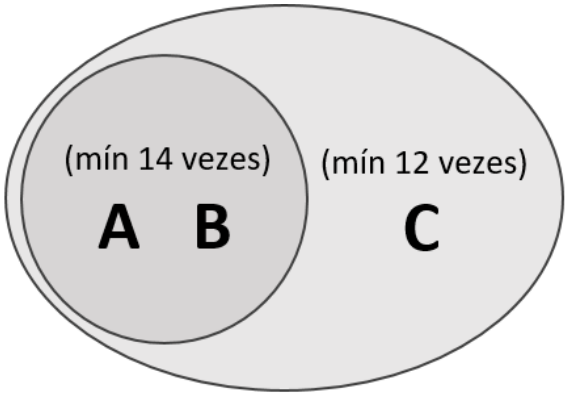

Figura 19 - Número mínimo de aparições do lado esquerdo e direito da regra
Fonte: Próprio autor.

Cabe destacar que esses valores de Suporte e Confiança podem ser alterados de acordo com a situação. Supondo que o auditor não encontre muitas suspeitas com 14 aparições conjuntas em licitações, ele pode reduzir o Suporte Mínimo de modo que as regras com frequência igual a 10 sejam consideradas na análise. De outra forma, caso o auditor tenha convicção de que o histórico de cartéis em sua região ocorra em torno de 20 licitações, ele também pode aumentar o valor do Suporte Mínimo a fim de podar mais o número de regras geradas. Analogamente, o valor da Confiança Mínima também pode ser alterado a critério dos auditores especialistas.

4.4.5 ALGORITMO DE SELEÇÃO DE REGRAS

Com o baixo valor de Suporte Mínimo adotado, é natural que haja a geração de um grande volume de Regras de Associação (RA), que neste caso foi de 26.821. Sendo assim, conforme explica Silva (2011), é necessário algum critério capaz de ranquear as melhores regras a fim de tornar viável a análise futura pela ação humana. Neste ponto, este artigo altera a metodologia original, propondo um novo Algoritmo de Seleção de Regras.

O algoritmo aqui proposto se baseia em quatro situações distintas, a fim de eliminar totalmente ou parcialmente regras que não atendam seus requisitos. Para exemplificar essas quatro situações, pode-se imaginar que as empresas “A”, “B”, “C” estão presentes no lado esquerdo da regra, com uma frequência de 15. Adicionalmente, conforme apresentado na figura 20, a empresa “D” figura no lado direito, com a frequência de 14.

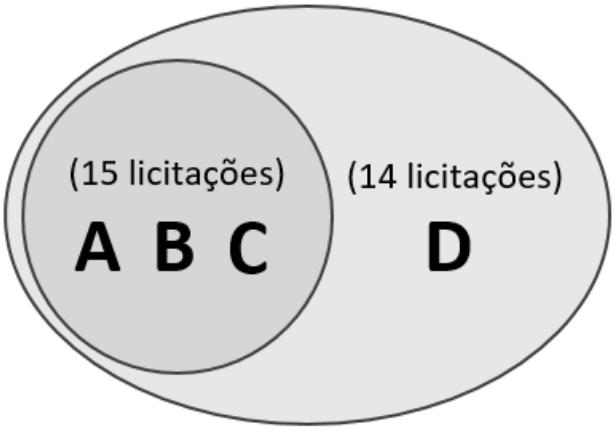

Figura 20 - Exemplo de regra de Associação
Fonte: Próprio autor.

A palavra-chave para entender o método criado chama-se Participação Relativa (PR), que varia de 0 a 100%, dada pela razão entre o número de participações em conjunto, e o número total de participações na base de dados, conforme equação: PR = 100x(Nº Participações Conjuntas / Nº Participações Total).

A título de exemplo, caso a empresa “A” tivesse participado de 140 licitações na base de dados, esta teria uma PR no valor de 10%. Em contrapartida, caso a mesma empresa tivesse participado de apenas 15 licitações, teria uma PR de 93%.

SITUAÇÃO 1: Quando nenhuma empresa tem PR alta dentro da regra. Isso significa que elas participam de muitas licitações, e acabam figurando a regra mesmo sem um número de editais significativo para a empresa, o que é muito comum quando se trata de grandes fornecedores. Neste caso, não há indícios da formação de cartel, e essas regras devem ser eliminadas.

SITUAÇÃO 2: Quando apenas uma empresa tem PR alta dentro da regra. Se o cartel é uma fraude praticada por um grupo de empresas, não há que se falar neste tipo de fraude perante esta situação. O que acontece é que uma empresa de menor porte acabou concorrendo com grandes fornecedoras em quase todas as suas participações, porém para as demais companhias, isso foi irrelevante. Sendo assim, essa regra também deve ser eliminada.

SITUAÇÃO 3: Quando duas ou mais empresas tem PR alta, mas não todas dentro da regra. Observe que para esse grupo, aquelas 14 licitações foram significativas, o que

evidencia um risco de cartel entre elas. Neste caso, deve-se eliminar empresas com PR baixo dentro da regra, e manter aquelas com alta PR dentro da suspeita para investigação mais apurada.

SITUAÇÃO 4: Todas as empresas têm PR alta dentro da regra. Perceba que neste caso, os 14 editais em conjunto significaram muito para todas as fornecedoras, sendo todas elas suspeitas de um conluio. Nessa situação, haverá o aproveitamento integral da regra.

Na figura 21 pode-se observar o fluxograma do Algoritmo de Seleção de Regras descrito acima, que elimina as regras das situações 1 e 2, aproveita parcialmente aquelas da situação 3, e aprova as da situação 4.

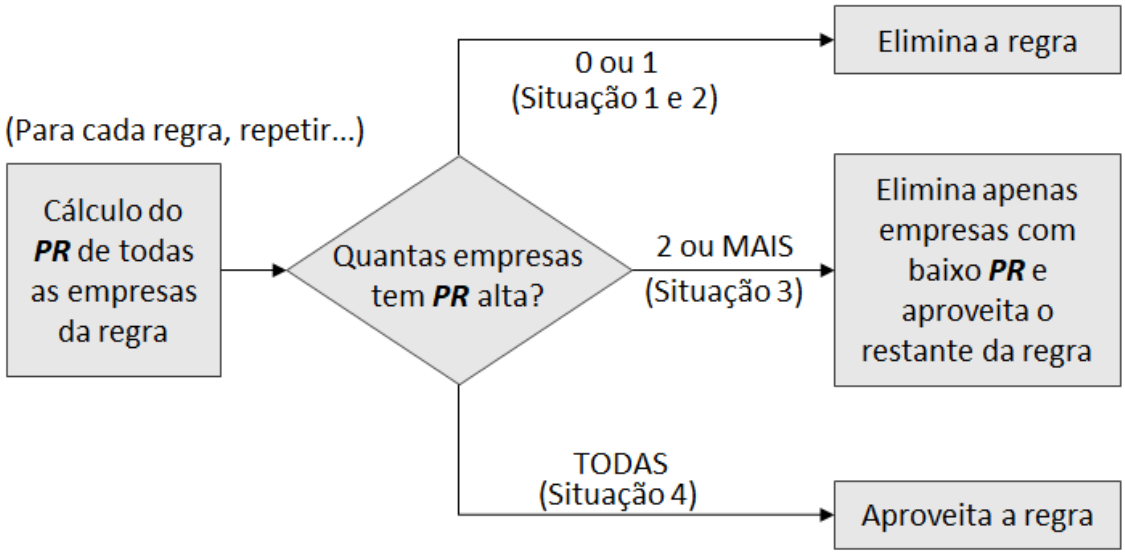

Figura 21 - Algoritmo proposto para seleção de regras

Fonte: Próprio autor.

Resta agora definir qual é o valor de corte da PR, de modo a classifica-la como alta ou baixa na aplicação do método. Neste estudo de caso foram feitos experimentos com diversos valores de Participação Relativa Mínima (PRM), onde o número de regras remanescentes após execução do algoritmo varia sensivelmente de acordo com PRM adotada. A tabela 11 e a figura 22 ilustram o que foi dito.

Cabe ressaltar que o objetivo deste artigo não é fixar um valor de PRM universal válido para qualquer situação. Na verdade, a proposta é deixar esse valor em aberto como um parâmetro de entrada, a ser variado caso a caso pelo especialista

encarregado da auditoria. Neste estudo, por exemplo, foi adotado um PRM de 50% inicialmente, e avaliadas as 11 regras remanescentes, como as principais suspeitas do estudo. Em seguida, foi adotado um PRM de 40% e avaliadas as 66 regras resultantes, que consequentemente apresentam indícios menos relevantes do que os anteriores.

Tabela 11 - Número de regras remanescentes de acordo com a PRM

Participação Relativa Mínima (PRM)	*Nº REGRAS*
0% (sem poda)	26.821
10%	25.210
20%	8.241
30%	748
40%	66
50%	11

Fonte: Próprio autor.

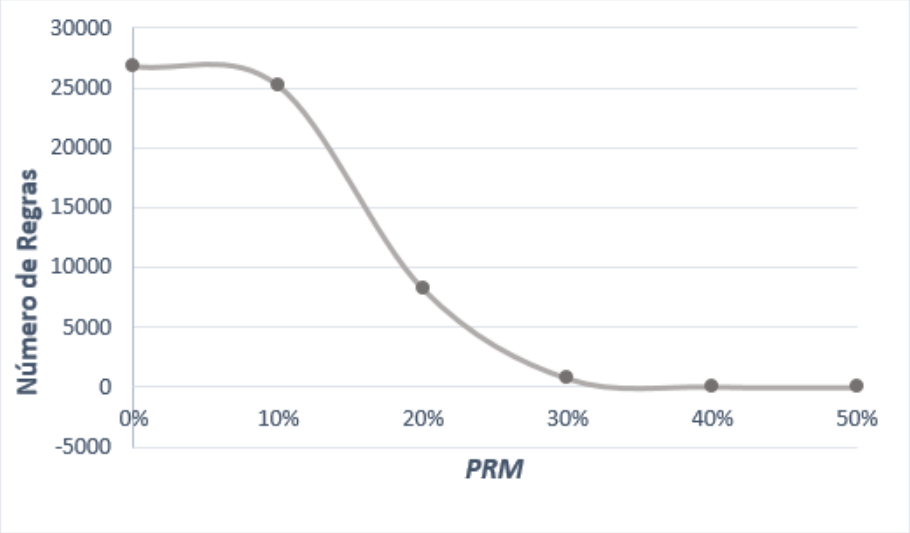

Figura 22 - Variação do número de regras remanescente com a PRM

Fonte: Próprio autor.

Em vista do acima exposto, fica claro que, quanto maior a PRM usada, menos regras são exibidas, e mais relevância elas têm. Cabe ao auditor começar por valores de PRM mais altos, e em seguida diminuí-lo a seu critério, até o ponto onde não julgar mais suspeito, tendo sempre em mente que o número de regras irá crescer à medida que a PRM diminui.

4.4.6 INTERPRETAÇÃO E AVALIAÇÃO DOS RESULTADOS

Após execução da Mineração de Dados (MD) por intermédio do algoritmo Apriori, com Suporte Mínimo de 0,1% e Confiança Mínima de 90%, conforme explicado no item 4.4.4, foram obtidas um total de 26.821 Regras de Associação (RA). Feito isso, essas regras foram alvo do Algoritmo de Seleção de Regras, proposto por este artigo no item 4.4.5, primeiramente com um PRM de 50%, gerando 11 regras, e em seguida com o PRM de 40%, gerando 66 regras remanescentes.

Como já mencionado, o CNPJ das empresas foi substituído por um número de identificação aleatório, devido ao caráter experimental do estudo, que visa apenas avaliar a metodologia, e não oferecer denúncia contra as companhias. Por esta razão, cabe deixar claro que não há nenhuma rastreabilidade entre o ID apresentado nas suspeitas, e o CNPJ real das empresas. Segue abaixo alguns resultados em caráter exemplificativo.

SUSPEITA 1: Um grupo formado por quatro fornecedores identificados pelos números 4982, 5681, 6167 e 3804 participou em conjunto de 18 licitações, com Participações Relativas (PR) acima de 50% para todas as empresas. Indo mais a fundo, percebeu-se que este grupo venceu dez, das 18 licitações disputadas, o que confere a estas empresas um forte indício de formação de cartel.

SUSPEITA 2: Um grupo formado por quatro fornecedores identificados pelos números 1634, 5023, 5082 e 2286 participou em conjunto de 42 licitações, porém com PR baixa para os dois últimos, o que os exclui da suspeita. Analisando melhor a relação entre os dois primeiros, percebe-se que as empresa 1634 e 5023 possuem PR de 82% e 51% respectivamente, o que indica chances razoáveis da formação de cartel entre elas. Apesar de o grupo ter vencido apenas sete licitações das 42 disputadas, o alto valor de PR da primeira chama a atenção para este possível conluio.

SUSPEITA 3: Um grupo formado por três fornecedores identificados pelos números 1721, 4600 e 3226 participou em conjunto de 19 licitações, com alta PR para os dois primeiros, e baixa para o último, excluído então da suspeita. Analisando as empresas 1721 e 4600, percebe-se que além dessas empresas terem vencido cinco das 19 licitações, elas apresentam PR de 53% e 51%, o que indica suspeita de cartel.

SUSPEITA 4: As empresas identificadas pelos números 1720 e 5023 participaram juntas de 47 licitações, o que representa uma PR de 68% e 53%, respectivamente. Apesar do elevado valor de licitações conjuntas, percebeu-se que este grupo não venceu nenhuma das referidas licitações, o que torna essa suspeita menos importante.

SUSPEITA 5: As empresas identificadas pelos números 3106 e 2599 participaram juntas de 19 licitações, o que representa uma PR de 76% e 63%, respectivamente. Este alto valor de PR, aliado ao fato de que este grupo venceu sete das 19 licitações, torna essa suspeita muito importante, trazendo à luz forte indício de formação de cartel entre esses dois fornecedores.

SUSPEITA 6: As empresas identificadas pelos números 1163 e 2437 participaram juntas de 24 licitações, o que se traduz em altos valores de PR, atingindo 92% para a primeira, e 69% para a segunda. Apesar de o grupo ter vencido apenas duas das 24 licitações, há forte indício da formação de cartel pelo altíssimo valor da Participação Relativa (PR) do fornecedor 1163.

4.5 CONCLUSÃO

A formação de cartéis por empresas privadas inidôneas visando lesar os cofres públicos é um problema vivido por governos de todo o mundo, como explicado na Introdução deste artigo. Neste contexto, surge como forte aliada nos processos de auditoria governamental, a Ciência de Dados, trazendo mais eficiência a essa fiscalização. É notório que os resultados obtidos por este artigo, que apontam empresas com mais chances de atuação ilícita, são de grande valia no planejamento dos auditores, que podem orientar seus esforços da maneira mais racional possível.

Sendo assim, este trabalho teve por primeiro objetivo descrever o arcabouço teórico referente à Descoberta de Conhecimento em Base de Dados (DCBD) aplicada no combate aos cartéis, e atingiu este resultado. Nesta ocasião, todas as etapas da DCBD foram descritas: Seleção, Pré-Processamento, Formatação, Mineração de Dados (MD) e Interpretação de Resultados. Vale destacar, que ainda no tópico 4.2 foi explicado o funcionamento do algoritmo utilizado neste estudo de caso, denominado

Apriori, capaz de gerar as Regras de Associação (RA) durante a etapa de MD. As estratégias inteligentes usadas por essa ferramenta para realizar as podas por Suporte Mínimo e Confiança Mínima foram exemplificadas, e o entendimento do algoritmo foi atingido.

Com relação à base de dados da CGU utilizada, conclui-se que esta é rica em detalhes, contemplando todas as licitações realizadas no âmbito da União, envolvendo os mais diversos objetos, de 2013 até os dias atuais. Fato que chama a atenção como ponto negativo no controle das contas públicas é a ausência de bancos de dados dos governos estaduais e municipais, que salvo algumas exceções, infelizmente não possuem dados disponíveis para estudo.

Sobre aspectos metodológicos, a primeira alteração com relação à proposta original, de Silva (2011), foi a eliminação da tarefa de *Clusterização*, que subdividia previamente o espaço amostral das licitações da União por regiões. Isso se justificou pelo fato de que o objeto escolhido tem por característica uma prestação que não se limita ao aspecto regional, já que podem ser prestadas à distância, sem limitações ou custos adicionais. Em vista dos bons resultados finais obtidos, pode-se concluir que esta alteração foi bem-sucedida, contudo deve-se sempre frisar que isto provavelmente não ocorrerá em casos de contratações com perfis mais regionais, sendo a *Clusterização* uma boa alternativa para melhorar os resultados nesses casos.

Sobre a etapa de Mineração de Dados (MD) também cabe algumas reflexões importantes. A primeira é que a tarefa de Associação se mostrou ideal na descoberta da relação entre as empresas que tendem a concorrer juntas, logo, conclui-se que esta Tarefa atendeu bem aos objetivos deste estudo. Outro ponto de discussão é sobre o parâmetro de entrada do algoritmo Apriori, Suporte Mínimo, que como explica Silva (2011), deve ter baixo valor a fim de não se perder regras valiosas. Isto é comprovado na análise de resultados, onde se pode observar que ótimas suspeitas possuem Suporte de apenas 0,1%. Por fim, conclui-se que a ferramenta WEKA, utilizada na MD, se mostrou bastante eficiente, de fácil operação e compreensão, e capaz de produzir resultados coerentes, de forma rápida, e de fácil interpretação.

Na etapa de Interpretação de resultados, tem-se o ponto mais importante deste artigo, pelo fato de que a metodologia original de Silva (2011) foi alterada, dando origem a um novo Algoritmo de Seleção de Regras. Esta ferramenta pós-mineração é essencial, pois o número de regras gerado é enorme, e atingiu mais de 26 mil neste estudo de caso. Como descrito no tópico 4.4.5, a palavra chave ao entendimento

dessa proposta é a Participação Relativa (PR), parâmetro que se mostrou eficiente na seleção de regras, e trouxe bons resultados finais. Cabe lembrar que este artigo não fixou a Participação Relativa Mínima (PRM), deixando esta como um *input* do algoritmo, a ser analisado caso-a-caso pelos auditores de acordo com a situação.

Por fim, conclui-se que as suspeitas levantadas neste artigo são de grande utilidade para que, durante o planejamento de uma auditoria governamental, essas empresas com indícios de fraude sejam auditadas de maneira prioritária, por apresentarem mais chances de formação de cartel. Note que tal recurso provavelmente tornará o processo mais assertivo, diminuindo assim a ocorrência deste conluio, e economizando recursos públicos com compras e contratações governamentais.

4.6 REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. *In:* INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994. **Proceedings** […]. Hove, East Sussex: Morgan Kaufmann, 1994. p. 487-499.

CAMILO, C.; SILVA, J. **Mineração de Dados:** Conceitos, Tarefas, Métodos e Ferramentas. Goiânia: Universidade Federal de Goiás, 2009.

CGU. **PORTAL TRANSPARÊNCIA,** 2019b. Disponível em: http://transparencia.gov.br/download-de-dados/licitacoes. Acesso em: 1 jul. 2019.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Advances in knowledge discovery and data mining**. Menlo Park, Calif.: AAAI Press, USA, 1996.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de Pesquisa**. Porto Alegre: UFRGS, 2009.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. São Paulo: Elsevier, 2005.

JUNG, C. F. **Metodologia Científica:** ênfase em Pesquisa Tecnológica. 3. ed. Porto Alegre: Penso, 2003.

LAKATOS, E. M.; MARCONI, M. DE A. **Metodologia científica.** 7. ed. São Paulo: Atlas, 2017.

LAROSE, D. T. **Discovering knowledge in data:** an introduction to data mining. Hoboken, N. J. : Wiley-Interscience, 2005.

NIEBUHR, J. de M.; NIEBUHR, P. de M. **Licitações e contratos das estatais**. Belo Horizonte: Fórum, 2018.

PIETRO, M. S. Z. D. **Direito Administrativo**. 32. ed. São Paulo: Forense, 2019.

SANTOS, F. B.; SOUZA, K. R. DE. **Como combater a corrupção em Licitações**. 2. ed. Belo Horizonte: Fórum, 2018.

SFERRA, H. H.; CORRÊA, A. C. J. Conceitos e Aplicações de Data Mining. **Revista de Ciência & Tecnologia,** Piracicaba, v. 11, n. 22, p. 19–34, 2003.

SILVA, C. V. S. **Agentes de Mineração e sua Aplicação no Domínio da Auditoria Governamental**. Brasília: Universidade de Brasília, 2011.

SUMATHI, S.; SIVANANDAM, S. N. **Introduction to data mining and its applications**. Berlin; New York: Springer, 2006.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

WAIKATO, U. **Weka 3:** Machine Learning Software in Java. 2019. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 21 out. 2019.

YIN, R. K. **Case study research:** design and methods. 5th. ed. Los Angeles: SAGE, 2014.

5 CONSIDERAÇÕES FINAIS

No Artigo A se deu o primeiro passo ao entendimento do assunto, onde a bibliometria realizada na base *Scopus Elsevier* trouxe à luz trabalhos interessantes, relacionados ao uso de inteligência computacional no combate à formação de cartéis. Nessa etapa o autor reuniu dissertações e artigos que serviram de base para o entendimento do atual “estado da arte”, o que foi essencial para o desenvolvimento da metodologia proposta. Como crítica a esta fase, pode-se dizer que a bibliometria foi realizada apenas na plataforma *Scopus Elsevier*, podendo ser mais completa se expandida a outras bases. Contudo, em defesa a esta dissertação, cabe destacar que foram estudados trabalhos de muitas outras fontes para realização deste estudo, apesar destes não estarem incluídos nas estatísticas do estudo bibliométrico.

Ainda no Artigo A, foi detalhado o referencial teórico em duas partes distintas. A primeira relativa às licitações públicas, oportunidade em que foram discutidos os conceitos básicos sobre este processo administrativo, bem como suas modalidades de execução, e seus tipos de fraude mais comuns. A segunda parte foi relativa à Descoberta de Conhecimento em Base de Dados (DCBD), onde cada etapa deste processo foi apresentada, com atenção especial para a Mineração de Dados (MD) aplicada por meio de Regras de Associação (RA). Nesta ocasião o autor pôde aprender sobre seus conceitos básicos, e foi mais a fundo, entendendo o funcionamento detalhado do algoritmo Apriori. Em vista disso, pode-se concluir que os objetivos do Artigo A foram todos atingidos com êxito.

No Artigo B, o objetivo central foi o de propor uma metodologia capaz de revelar indícios da formação de cartéis. Tal metodologia foi baseada fortemente no trabalho de Silva (2011), com duas alterações importantes. A primeira foi a eliminação da tarefa de *Clusterização* quando se trata de um objeto licitatório cuja prestação tem caráter não regionalizado. Já a segunda mudança foi a proposta de um novo Algoritmo de Seleção de Regras, ferramenta essencial, visto que a etapa de Mineração de Dados (MD) produz um enorme volume de regras. Nesta ocasião, o algoritmo foi explicado com exemplos práticos envolvendo empresas fictícias para melhor entendimento do mesmo. Sendo assim, conclui-se que este artigo atingiu seus objetivos, e trouxe pleno entendimento sobre a proposta no novo método.

Por fim, o Artigo C teve como objetivo testar a metodologia proposta no Artigo B com uma base de dados real de licitações, obtida no portal da CGU, a fim de se

verificar a validade e eficiência do método. O primeiro ponto a ser discutido é sobre a eliminação da tarefa de *Clusterização*, que a julgar pelos bons resultados obtidos, foi uma alteração bem-sucedida no método original. Contudo, como já explicado no referido artigo, cabe frisar que tal alteração só é indicada para objetos cuja prestação pode ser feita à distância sem custos adicionais, livrando assim este mercado de qualquer interferência regional.

Sobre a etapa de Mineração de Dados (MD), seguiu-se o método original de Silva (2011), e pelos bons resultados observados no Artigo C, comprovou-se que as chamadas Regras de Associação (RA) com o algoritmo Apriori são uma excelente alternativa para solução deste problema. É interessante comentar também sobre a constatação de que o parâmetro de entrada Suporte Mínimo deve ter seu valor baixo, visto que muitas regras relevantes encontradas possuem baixíssimo valor de Suporte, como já afirmava Silva (2011).

Ainda sobre o Artigo C, o ponto mais importante foi a validação do novo Algoritmo de Seleção de Regras, que foi proposto no Artigo B, a fim de se eliminar regras com menor qualidade, reduzindo assim o volume de informações. Sobre ele, cabe destacar que seu parâmetro de entrada, Participação Relativa Mínima (PRM), fica para livre escolha do usuário, e que quanto maior é esse valor, menor é o número de regras remanescentes para análise humana. Ponto muito positivo é a flexibilidade deste algoritmo, que dá liberdade ao auditor para testar a PRM adequada a cada situação, calibrando assim o número de suspeitas a serem investigadas, de acordo com o tempo e recursos disponíveis.

Por fim, analisando o trabalho como um todo perante os objetivos gerais e específicos traçados no tópico 1.2, conclui-se que este foi bem-sucedido, e cumpriu o que se propôs a fazer. Contudo, como todo trabalho acadêmico, há pontos de melhoria e expansão deste estudo que podem ser implementados futuramente. A principal é com relação ao Algoritmo de Seleção de Regras, que se baseia somente na PRM para avaliar a qualidade das suspeitas, sendo que outros parâmetros também poderiam ser avaliados simultaneamente com aquele, tais como o número de licitações vencidas pelo grupo suspeito.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. *In:* INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994. **Proceedings** […]. Hove, East Sussex: Morgan Kaufmann, 1994. p. 487-499.

BAJARI, P.; YE, L. **Deciding Between Competition and Collusion.** Review of Economics and Statistics, Massachusetts, v. 85, n. 4, p. 971–989, Massachusetts, nov. 2003.

CAMILO, C.; SILVA, J. **Mineração de Dados:** Conceitos, Tarefas, Métodos e Ferramentas. Goiânia: Universidade Federal de Goiás, 2009.

CGU. **Observatório da Despesa Pública participa de congresso internacional sobre mineração de dados**. 2016. Disponível em: https://www.cgu.gov.br/noticias/2016/08/observatorio-da-despesa-publica-participa-de-congresso-internacional-sobre mineracao-de-dados. Acesso em: 5 ago. 2019.

CGU. **Observatório da Despesa Pública.** 2019a. Disponível em: https://www.cgu.gov.br/assuntos/informacoes-estrategicas/observatorio-da-despesa - publica. Acesso em: 5 ago. 2019.

CGU. **PORTAL TRANSPARÊNCIA,** 2019b. Disponível em: http://transparencia.gov .br/download-de-dados/licitacoes. Acesso em: 1 jul. 2019.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Advances in knowledge discovery and data mining**. Menlo Park, Calif.: AAAI Press, USA, 1996.

GABARDO, A. C.; LOPES, H. S. Using Social Network Analysis to Unveil Cartels in Public Bids. *In*: EUROPEAN NETWORK INTELLIGENCE CONFERENCE (ENIC), 14., 2014, Wroclaw, Poland. **Anais** […]. Washington: IEEE, 2014. p. 17-21.

GANTZ, J.; REINSEL, D. **The Digital Universe In 2020:** Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. IDC Corporate, Framingham, dec. 2012. Disponível em: https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf. Acesso em: 14 jun. 2019.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de Pesquisa**. Porto Alegre: UFRGS, 2009.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. São Paulo: Elsevier, 2005.

IBPAD. **Ciência de Dados no Combate à Corrupção**. 2016. Disponível em: https://www.ibpad.com.br/blog/analise-de-dados/ciencia-de-dados-no-combate-corrupcao/. Acesso em: 5 ago. 2019.

JUNG, C. F. **Metodologia Científica:** Ênfase em Pesquisa Tecnológica. 3. ed. Porto Alegre: Penso, 2003.

LAKATOS, E. M.; MARCONI, M. DE A. **Metodologia científica.** 7. ed. São Paulo: Grupo Gen - Atlas, 2017.

LAROSE, D. T. **Discovering knowledge in data:** an introduction to data mining. Hoboken, N. J.: Wiley-Interscience, 2005.

LICHTBLAU, E. **F.B.I. Data Mining Reached Beyond Initial Targets**. The New York Times, Washington, 9 set. 2007. Disponível em: https://www.nytimes.com/2007/09/ 09/washington/09fbi.html. Acesso em: 14 jun. 2019.

MAJADI, N.; TREVATHAN, J.; BERGMANN, N. Real-Time Collusive Shill Bidding Detection in Online Auctions. *In:* MITROVIC, T.; XUE, B.; LI, X. (eds.). ***AI 2018:*** Advances in Artificial Intelligence. Cham: Springer International Publishing, 2018. v. 11320. p. 184–192.

NIEBUHR, J. de M.; NIEBUHR, P. de M. **Licitações e contratos das estatais**. Belo Horizonte: Fórum, 2018.

PADHI, S. S.; MOHAPATRA, P. K. J. Detection of collusion in government procurement auctions. **Journal of Purchasing and Supply Management**, London, v. 17, n. 4, p. 207–221, dez. 2011.

PIETRO, M. S. Z. D. **Direito Administrativo**. 32. ed. São Paulo: Forense, 2019.

PIXININE, J. **Na memória:** relembre a evolução dos dispositivos de armazenamento. 2017. Disponível em: https://www.techtudo.com.br/listas/noticia/ 2015/05/na-memoria-relembre-a-evolucao-dos-dispositivos-de-armazenamento.html. Acesso em: 8 maio. 2019.

SANCHEZ-GRAELLS, A. 'Screening for Cartels' in Public Procurement: Cheating at Solitaire to Sell Fool's Gold? **Journal of European Competition Law & Practice**, Oxford, v. 10, n. 4, p. 199–211, 1 abr. 2019.

SANTOS, F. B.; SOUZA, K. R. DE. **Como combater a corrupção em Licitações**. 2. ed. Belo Horizonte: Fórum, 2018.

SFERRA, H. H.; CORRÊA, A. C. J. Conceitos e Aplicações de Data Mining. **Revista de Ciência & Tecnologia,** Piracicaba, v. 11, n. 22, p. 19–34, 2003.

SILVA, C. V. S. **Agentes de Mineração e sua Aplicação no Domínio da Auditoria Governamental**. Brasília: Universidade de Brasília, 2011.

SILVA, J. A. **Curso de Direito Constitucional Positivo.** 42. ed. São Paulo: Malheiros, 2019.

SUMATHI, S.; SIVANANDAM, S. N. **Introduction to data mining and its applications**. Berlin; New York: Springer, 2006.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

WAIKATO, U. **Weka 3:** Machine Learning Software in Java. 2019. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 21 out. 2019.

YIN, R. K. **Case study research: design and methods**. Fifth edition ed. Los Angeles: SAGE, 2014.