Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Paulo Jhonny Scheleder da Costa Rosa

Descartelizando: Uso de *Machine Learning* e Estatística para Detecção de Indícios de Cartel em Processos Licitatórios

Curitiba
2019

Paulo Jhonny Scheleder da Costa Rosa

Descartelizando: Uso de *Machine Learning* e Estatística para Detecção de Indícios de Cartel em Processos Licitatórios

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Walmes Marques Zeviani

Curitiba
2019

Descartelizando: Uso de *Machine Learning* e Estatística para Detecção de Indícios de Cartel em Processos Licitatórios

Descartelizando: Use of Machine Learning and Statistics for Detection of Cartel Signs in Bidding Processes

Paulo Jhonny Scheleder da Costa Rosa[1]

[1]Ministério Público do Estado do Paraná, Rua Mauá, 920, Curitiba, PR, Brasil*

Resumo

Propõe-se, neste artigo, uma metodologia para detecção de indícios de cartel em licitações relativas à área de engenharia, considerando a gestão municipal paranaense de 2013 a 2016. A partir de dados públicos e de técnicas estatísticas e de *machine learning*, buscou-se identificar mercados de licitação, encontrar regras de associação mediante a atuação conjunta e frequente de empresas e criar um indicador de risco de cartel. Os resultados apontaram três mercados de licitação compostos por regiões vizinhas, identificados via técnica k-means. Além disso, considerando o algoritmo *Apriori* e o estado do Paraná, 245 potenciais carteis foram encontrados e avaliados com base em seu sucesso contratual. Cerca de 5,4% do total de licitações tiveram a aplicação de pelo menos uma regra, perfazendo uma média de risco de cartel de 0,838. Por fim, a partir da rede neural SOM, observou-se uma associação negativa entre o risco de cartel e o número de participantes da licitação, indicando que ambientes com poucos competidores contribuem para atuações colusivas.
Palavras-chave: licitação, cartel, k-means, regras de associação, rede neural SOM, indicadores de risco

Abstract

This article proposes a methodology for detecting signs of cartel in bidding processes related to engineering area, considering the municipal management of Paraná from 2013 to 2016. Based on public data and statistical and machine learning techniques, we sought to identify bidding markets, find association rules through joint and frequent performance of companies and create a indicator of cartel risk. The results pointed out three bidding markets composed of neighboring regions, identified through the k-means technique. In addition, considering the Apriori algorithm and the state of Paraná, 245 potential cartels were found and evaluated based on their contractual success. About 5.4% of the total biddings had the application of at least one rule, resulting in an average cartel risk of 0.838. Finally, from the SOM neural network, a negative association was observed between the cartel risk and the number of companies in the bidding, indicating that environments with few competitors contribute to collusive actions.
Keywords: bidding, cartel, k-means, association rules, SOM neural network, risk indicators

1. Introdução

Segundo a Lei Federal 8.666/93 [1], um processo licitatório é o meio administrativo pelo qual a Administração Pública adquire bens, obras e serviços indispensáveis ao cumprimento de suas obrigações, tendo como objetivo escolher, dentre vários competidores, a proposta mais vantajosa no que concerne aos aspectos de preço e qualidade. Por esse motivo, é imprescindível a existência de competição entre os participantes e a escolha da melhor proposta, premissas fundamentais vinculadas aos princípios de igualdade e legalidade regidos por aquela Lei. Em paralelo, um contrato administrativo é todo e qualquer ajuste entre a Administração Pública e particulares, em que haja um acordo de vontades para a formação de vínculo e a estipulação de obrigações recíprocas, tal como disposto no art. 2º, parágrafo único, da Lei Federal 8.666/93. Além disso, segundo o art. 54º, inciso 1º, desta mesma lei, os contratos devem estabelecer com clareza e precisão as condições para sua execução, expressas em cláusulas que definam os direitos, obrigações e responsabilidades das partes, em conformidade com os termos da licitação e da proposta a que se vinculam.

*pjsdcrosa@mppr.mp.br

No entanto, em virtude do grande volume de recursos públicos envolvidos, os princípios que regem tanto processos licitatórios quanto contratos públicos podem ser infringidos mediante a prática de atividades ilícitas, responsáveis por lesar o caráter competitivo entre os participantes. Em geral, empresas integrantes de esquemas cooperativos visam à participação em certames que envolvam o pagamento de valores monetários expressivos, como é o caso de processos de compra para aquisição de produtos, obras e serviços de engenharia. Conforme pesquisa realizada por Carazza (2016) [2], cerca de 30% das doações para campanhas políticas do ano de 2014 foram oriundas de empresas do setor econômico de construção. O autor explica que o foco dessas empresas, ao doar quantias volumosas para candidatos em época eleitoral, seria o de obter um retorno do dinheiro doado, quer seja por meio de contratações futuras milionárias ou qualquer outra forma de recebimento.

Dentre as práticas anticoncorrenciais utilizadas está a formação de cartel e o rodízio entre vencedores. Um cartel é uma espécie de acordo entre empresas no sentido de combinar preços para eliminar a concorrência, podendo haver revezamento entre os participantes do esquema na contratação com a Administração Pública. Segundo a Cartilha *Combate a Carteis em Licitações* [3], elaborada pelo Ministério da Justiça, estimativas da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) demonstram que carteis geram um sobrepreço estimado entre 10% e 20% comparado ao preço de um mercado competitivo.

Ishii (2009) [4] explica que a atuação de carteis implica em contratos com valor relativamente mais alto e próximo de um valor de referência orçado pela Administração Pública, haja vista que empresas participantes de conluio dificilmente propõem preços inferiores a 90% ou 95% do valor estimado. Conforme as *Diretrizes para Combater o Conluio entre Concorrentes em Contratações Públicas* [5], estabelecidas pela OCDE, esquemas de cartel em licitações frequentemente incluem mecanismos de partilha dos lucros adicionais obtidos por meio da contratação por preço final mais elevado. Os concorrentes que combinam de abandonar o certame ou apresentar propostas para perder podem ser subcontratados pelo concorrente cuja proposta foi adjudicada, de forma a dividir os lucros obtidos a partir da proposta com o preço mais elevado, alcançados de forma ilegal. Contudo, os carteis em licitações podem utilizar métodos muito mais elaborados para obtenção de adjudicações de contratos e divisão dos lucros.

No Brasil, o Conselho Administrativo de Defesa Econômica (CADE) é responsável por investigar e punir empresas que se unem na formação de cartel, prática que configura tanto ilícito administrativo punível pelo CADE, nos termos da Lei nº 8.884/94 [6], quanto crime punível com pena de 2 a 5 anos de reclusão, nos termos da Lei nº 8.137/90 [7]. Ademais, em 2009, a Secretaria de Direito Econômico firmou acordos de cooperação com a Controladoria-Geral da União (CGU) e o Tribunal de Contas da União (TCU) no intuito de somar esforços junto a acordos já existentes entre a Polícia Federal e os Ministérios Públicos Federal e Estaduais para combater a prática de corrupção e lavagem de dinheiro, crimes geralmente realizados em conjunto com atividades ilícitas em processos licitatórios (SILVA, 2011) [8].

O Ministério Público do Estado do Paraná (MPPR), por meio das Promotorias de Justiça do Patrimônio Público, tem uma atuação recorrente em relação ao combate a fraudes em licitações públicas. No ciclo político de 2013 a 2016, foram registrados cerca de sete mil procedimentos para investigar práticas ilícitas em processos licitatórios[1]. No entanto, na maioria das vezes, as investigações são iniciadas a partir de denúncias anônimas realizadas por terceiros, geralmente empresas prejudicadas pela existência de conduta anticompetitiva no certame do qual participaram. No mais, o crescimento exponencial do volume de contratações públicas municipais em um curto espaço de tempo e a inexistência de mecanismos ágeis para detecção de indícios de práticas anticoncorrenciais impõem desafios à atuação proativa dos guardiões da lei. Por conseguinte, torna-se imperiosa a implementação de metodologias científicas que permitam a produção de conhecimento pautado à formação de convicção fundamentada, visando subsidiar tomadores de decisão no enfrentamento dessas práticas de forma eficiente e consistente.

Na literatura internacional, é possível encontrar pesquisas científicas voltadas à identificação de padrões estatísticos relacionados à formação de cartel em procedimentos licitatórios. Ishii (2009) [4] utilizou-se de modelos econométricos para estudar o preço pago em relação ao estimado em licitações realizadas em Naha, no Japão. O autor constatou que é possível caracterizar indícios de cartel quando o preço contratado se

[1] PROMP, MPPR, 22/04/2019 - considerou-se as seguintes palavras-chave: contratação irregular, contrato, contrato e serviços públicos, dispensa de licitação, inexigibilidade de licitação, licitação, licitação (antigo), procedimento licitatório e superfaturamento.

aproxima sobremaneira do preço orçado, tendo como ponto de corte o percentual de 95%. Padhi (2011) [9] estudou a distribuição da razão entre os preços propostos por empresas licitantes comparados aos valores orçados pela Administração Pública da Índia. A partir da aplicação de técnicas de clusterização, foi possível dividir a referida razão em dois grupos estatisticamente diferentes: o primeiro, por apresentar média e variância altas, foi caracterizado como um ambiente colusivo, ao passo que o segundo, por não apresentar tais características, foi considerado como um conjunto de certames em que o princípio de competitividade não foi violado. Morozov (2013) [10], por sua vez, ajustou modelos de regressão para explicar a variabilidade do percentual pago pela Administração Pública russa a partir de covariáveis relativas às empresas contratadas. Os resultados demonstraram que a experiência da empresa, no que diz respeito a contratações passadas e sua capacidade operacional para realização do serviço, não tem efeito no preço contratado.

No que se refere à literatura nacional, Silva (2011) [8] investigou possíveis conluios em procedimentos licitatórios realizados pelo Governo Federal a partir do uso de técnicas de mineração de dados e sistemas multiagentes. O autor desenvolveu uma arquitetura denominada *AGent Mining Integration* (AGMI) por meio da integração de diferentes técnicas de mineração de dados junto a uma abordagem multiagentes para automatização do processo de descoberta de conhecimento. Considerando somente o uso da técnica de regras de associação, os experimentos com a AGMI mostraram um aumento de 170% na qualidade média das dez melhores regras encontradas, auxiliando de forma consistente a detecção de carteis. Morais (2016) [11] propôs uma série de indicadores para investigação de atividades ilícitas em procedimentos licitatórios, cuja construção foi baseada em características de empresas participantes, órgãos licitantes, licitações e contratos. Fraga (2017) [12] buscou por indícios de colusão entre participantes de licitações municipais realizadas pelo estado da Paraíba no período de 2005 a 2016. A partir da aplicação da técnica de regras de associação, foram encontrados fortes indícios de suspeição de conluio para várias empresas dos ramos de alimentação, prestação de serviços de limpeza, etc.

Diante do exposto, busca-se desenvolver neste estudo uma metodologia para detecção de indícios de cartel em processos licitatórios relativos à aquisição de produtos, obras e serviços de engenharia, considerando a gestão municipal paranaense de 2013 a 2016. Espera-se que os resultados obtidos auxiliem a atividade finalística do MPPR na atuação proativa a respeito de possíveis irregularidades envolvendo contratações públicas, garantindo a defesa do Patrimônio Público e o combate à corrupção.

2. Materiais e Métodos

Inicialmente, a partir de dados extraídos de fontes públicas, faz-se uso de técnicas de clusterização para identificar mercados de licitação, mapear a atuação de empresas e particionar o espaço de soluções. Para cada uma das divisões encontradas, aplica-se o algoritmo *Apriori* de descoberta de regras de associação a fim de detectar potenciais carteis caracterizados pela atuação frequente e conjunta de empresas.

Na sequência, faz-se uma extensão de funções de avaliação de regras de associação propostas por Silva (2011) [8] e Fraga (2017) [12], no sentido de se obter um resultado mais consistente em relação a sua qualidade e assertividade na caracterização de indícios de cartel. Nesse sentido, são propostas duas funções de avaliação, quais sejam: a probabilidade de o grupo de empresas firmar um contrato com a Administração Pública e a mediana das razões entre os valores dos contratos celebrados pelo grupo de empresas e os valores orçados pela Administração Pública. Os resultados são representados por um indicador numérico de risco de cartel.

Por fim, os procedimentos licitatórios são classificados com base no risco de cartel associado. As soluções encontradas são validadas por meio de uma análise associativa entre o indicador criado e fatores de risco estabelecidos a partir de características de licitações, contratos e empresas. Para tanto, faz-se uso de uma rede neural artificial com aprendizado não supervisionado, denominada *Self-Organizing Maps* (SOM).

2.1. Bases de Dados

- **Licitações e Contratos**

No Paraná, os municípios devem repassar dados de licitações e contratos ao Tribunal de Contas do Estado do Paraná (TCE/PR), órgão que tem como uma de suas principais funções fiscalizar atividades administrativas de Entidades da Administração Pública Municipal. Nos termos do art. 29º da Lei Complementar nº 113 de 15/12/2005 [13], para assegurar a eficácia do controle e instruir o julgamento das contas, o Tribunal efetua a fiscalização dos atos praticados por entes sujeitos

à sua jurisdição que resultem em receita ou despesa, sendo uma de suas responsabilidades acompanhar, pela publicação na imprensa oficial ou por outro meio, os editais de licitação, os contratos, inclusive administrativos, e os convênios, acordos, ajustes ou outros instrumentos congêneres.

Os dados fornecidos pelas entidades municipais estão disponíveis no Portal de Informações para Todos (PIT) [14] do TCE/PR, instrumento de consulta criado em 2016 no qual é possível acessar dados relativos a licitações públicas, contratos, convênios, obras, despesas, combustíveis e diárias referentes aos 399 municípios paranaenses. O referido portal oferece também a possibilidade de extrair dados em arquivos *XML* (*eXtensible Markup Language*), que é uma linguagem de marcação recomendada para a criação de documentos com dados organizados hierarquicamente, em contraposição aos formatos de dados tabulares (*csv*, por exemplo), tais como textos, banco de dados semiestruturados e desenhos vetoriais.

Os arquivos *XML* são disponibilizados no painel Dados Abertos, o qual armazena dados brutos em formato aberto para *download* por município e ano. Os dados de 323.642 licitações e 380.863 contratos realizados no período de 2013 a 2016 foram extraídos do aludido painel via técnicas de *web scraping*. As rotinas de programação foram implementadas por meio do *software R 3.4.1* [15] na biblioteca denominada *LicitaR*, que inclusive está disponível no *Github* [16] para acesso da população. O referido pacote apresenta funções específicas para a coleta dos arquivos *XML*, bem como para seu tratamento e formatação em estrutura *tidy*, ou seja, linhas correspondendo aos registros e colunas aos atributos.

- **Dados de Empresas Licitantes**

Os dados cadastrais, quadros societários e as atividades econômicas das empresas participantes de licitação no período de 2013 a 2016 foram obtidos por meio de uma função, implementada na biblioteca *LicitaR*, que captura dados a partir da *API*[2] (*Application Programming Interface*) denominada Receita WS [17], criada para recuperação de informações de empresas. A inidoneidade das empresas até o ano de 2017, um ano após o término do ciclo político de 2013 a 2016, foi verificada a partir do Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) [18] mantido pela CGU. Por fim, buscou-se por informações a respeito de doações de campanha para as Eleições Municipais de 2012, por meio do Repositório de Dados Eleitorais [19] do Tribunal Superior Eleitoral (TSE).

[2] conjunto de rotinas e padrões de programação para acesso a um aplicativo de *software* ou plataforma baseado na *web*.

2.2. Preparação dos Dados

Os processos licitatórios relacionados à aquisição de produtos, obras e serviços de engenharia foram selecionados considerando um dicionário de palavras-chave composto por termos correlatos ao assunto. A seleção das palavras se deu por meio da aplicação de técnicas de mineração de texto para limpeza, formatação e seleção dos objetos de licitação. Tal medida foi necessária pelo fato de os dados disponíveis no PIT não estarem categorizados por temáticas, tornando inviável a seleção via mecanismos de filtro. Desta forma, foram selecionadas 26.027 licitações cuja descrição do objeto apresentou pelo menos uma das palavras-chave do dicionário temático.

Além disso, 858 registros foram retirados da base de dados por apresentarem inconsistências consonante a alguns critérios estabelecidos à priori, quais sejam: licitações nas modalidades dispensa ou inexigibilidade, por não apresentarem concorrência; licitações sem participantes; procedimentos licitatórios em que pelo menos um dos participantes se enquadrou como pessoa física; licitações com valor de referência orçado pela Administração Pública igual a zero; licitações cujos contratos apresentaram valor igual a zero ou a soma ultrapassou o valor de referência; licitações cuja data de assinatura do contrato foi cadastrada como anterior à data do edital; licitações com valor de referência menor que R$ 1.000 e maior que R$ 20.000.000. Os *datasets* de licitações e contratos após os procedimentos supracitados passaram a conter 25.169 e 34.512 registros, respectivamente.

2.3. Análise Espacial e Clusterização

Nesta etapa, utilizou-se de algoritmos de clusterização para identificar possíveis mercados de licitação e dividir o espaço de soluções, considerando a participação de empresas em processos licitatórios por gepatrias, que são unidades do MPPR responsáveis pela atuação na área do Patrimônio Público, divididas territorialmente em 11 regiões disjuntas do Paraná.

2.3.1. Mapas de Fluxo Origem-Destino

Primeiramente, foi realizada uma análise espacial a partir da identificação de fluxos origem-destino de participação de uma parcela de empresas nos gepatrias. Decidiu-se por fazê-la a fim de entender a abrangência municipal de atuação das empresas e preparar o *dataset* a ser utilizado como entrada para os algoritmos. Entretanto, pelo fato de não ser possível representar todos os fluxos entre as 10.813 empresas e os 11 gepatrias, de forma interpretável e amigável em um mapa, foram selecionadas apenas empresas localizadas no Paraná com um número de participações em processos licitatórios igual ou superior a 20 (658 empresas), assumindo-se que empresas detentoras de menos de 20 participações seguiriam uma tendência de fluxo parecida, tendo possivelmente uma abrangência de participação compreendida num raio próximo de sua localização geográfica. Outrossim, ao retirar empresas que participaram poucas vezes e em apenas um gepatria, buscou-se garantir a minimização da falta de informação.

Na sequência, foi realizada a leitura e preparação, no *R*, de um *shapefile* relativo aos municípios do Paraná, disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) [20]. Inicialmente, foram retirados dois polígonos relativos aos municípios de Nova Cantu e Rio Bom, cuja quantidade de licitações para aquisição de produtos, obras e serviços de engenharia foi nula no período. Ainda, como as unidades jurisdicionais do MPPR para atuação na área do Patrimônio Público são representadas por gepatrias, que são conglomerados de municípios vizinhos geograficamente, foi necessário utilizar funções do *R* para unir polígonos de municípios pertencentes a um mesmo gepatria em um novo polígono. Ademais, foram adicionadas à tabela de atributos do *shapefile* variáveis necessárias para criação de um mapa de fluxo, como a latitude e longitude relacionada não só à localização de cada empresa como também ao centroide de cada gepatria e a quantidade de participações das empresas nos respectivos gepatrias.

2.3.2. Algoritmos de Clusterização

No que tange aos algoritmos de clusterização, utilizou-se das técnicas SKATER, k-means e métodos hierárquicos aglomerativos.

- **SKATER**

É uma técnica de clusterização espacial criada por Assuncao (2006) [21] e implementada no pacote *spdep* do *software R*. O algoritmo é baseado na poda de uma árvore geradora mínima construída a partir da estrutura de vizinhança das unidades espaciais a serem agrupadas. Os grupos obtidos devem ser internamente homogêneos com relação a atributos de interesse e, ao mesmo tempo, devem ser heterogêneos entre si. Considerou-se como entrada para o algoritmo o *shapefile* de gepatrias, cuja tabela de atributos apresentou, como variáveis de discriminação, as quantidades de participações de cada uma das 658 empresas nos 11 gepatrias. Para evitar distorções na estrutura dos agrupamentos, todas as variáveis foram padronizadas segundo a estatística *z-escore*, dada pela expressão $\tilde{z} = \frac{(x-\bar{x})}{\hat{\sigma}_x}$. Além disso, o parâmetro de poda foi setado no conjunto $c = (1,2)$ a fim de encontrar 2 e 3 *clusters*, respectivamente, e a medida de distância entre os vetores de atributos escolhida foi a euclidiana.

- **k-means e métodos hierárquicos**

Segundo Fávero (2009) [22], o método não hierárquico k-means tem como objetivo encontrar uma partição de n elementos em k grupos (*clusters*), de modo que a partição satisfaça dois requisitos básicos: semelhança interna e separação dos *clusters* formados. Para tanto, o processo de treinamento é composto por três passos, quais sejam: 1) Partição inicial dos indivíduos em k *clusters*, cuja definição deve ser feita pelo analista; 2) Cálculo dos centroides para cada um dos k *clusters* e cálculo da distância euclidiana dos centroides em relação a cada registro da base de dados; e 3) Agrupar os registros aos *clusters* cujos centroides se encontram mais próximos, e voltar ao passo 2 até que não ocorra variação significativa na distância mínima de cada registro da base de dados em relação a cada um dos centroides dos k *clusters*.

Os métodos hierárquicos, por sua vez, podem ser divididos em dois tipos de agrupamento: aglomerativos e divisivos. No método aglomerativo, cada registro começa com seu próprio agrupamento e, a partir deste ponto, novos agrupamentos são realizados por similaridade. Já no método divisivo, todas as observações começam em um grande agregado, sendo separadas primeiramente as observações mais distantes, até que cada observação se torne um grupo isolado. Neste estudo, foram utilizados os seguintes algoritmos hierárquicos aglomerativos: Ligação Individual (*single*), Li-

gação Completa (*complete*), Ligação Média (*average*), Centroide, Mediana e *Ward*.

Sendo assim, a partir dos métodos supracitados, buscou-se clusterizar as empresas licitantes no Paraná com base na contagem de participações em cada um dos 11 gepatrias, ou seja, o *dataset* de entrada para os algoritmos foi configurado de tal forma que as linhas correspondessem às empresas e as colunas aos gepatrias. O número ótimo de partições foi definido por meio da biblioteca *NbClust* do *R*, que fornece cerca de 30 índices para determinar a quantidade ideal de *clusters*. Além disso, todas as variáveis foram também padronizadas segundo a estatística *z-escore* e uma normalização por linhas, em que foi considerada a proporção de participação de cada empresa nos 11 gepatrias aos invés da contagem bruta, dada por $p(i,j) = \frac{contagem(i,j)}{\sum_{j=1}^{11} contagem_{ij}}$, sendo $i = 1,2,3,...,658$ empresas e $j = 1,2,...,11$ gepatrias. A medida de distância entre os vetores de atributos escolhida foi a euclidiana.

Após a obtenção dos grupos de empresas, foi realizada uma Análise de Variância (ANOVA), a partir do teste *F*, para comparar as médias de participação em relação a cada um dos 11 gepatrias. Na sequência, cada gepatria foi associado ao grupo com maior proporção de participação média das empresas, formando-se, então, *clusters* de gepatrias.

2.4. Descoberta de Regras de Associação

2.4.1. Apresentação e Procedimentos Adotados

Segundo Goldschmidt (2015) [23], a descoberta de regras de associação é uma técnica de mineração de dados que tem por objetivo encontrar itens que implicam na presença de outros itens em uma mesma transação, detectando padrões em forma de regras. Agrawal (1994) [24] explica que uma regra de associação fornece uma relação entre atributos de uma base de transações. Seja D uma base de transações e $I = I_1, I_2, ..., I_m$ um conjunto de m itens distintos de D, em que cada transação T tem um conjunto de itens de tal modo que $T \subseteq I$ e tem um identificador único. Uma transação T contém um conjunto de itens X se, e somente se, $X \subseteq T$.

Uma regra de associação é uma implicação da forma $X \Rightarrow Y$, em que $X \subset I$, $Y \subset I$ e $X \cap Y = \emptyset$. Os *itemsets* X e Y são chamados de subsequente e consequente. A regra $X \Rightarrow Y$ pertence à base de transações D com suporte (s), em que s é a razão do total de registros que contêm $X \cup Y$ (ou seja, ambos os conjuntos X e Y) pelo total de registros da base de dados. A regra $X \Rightarrow Y$ tem confiança (c), que é a razão do número de registros que contêm $X \cup Y$ pelo número de registros que contêm apenas X. O problema geralmente é decomposto em duas partes: 1) Encontrar todos os conjuntos de itens que ocorrem com uma frequência maior ou igual ao suporte mínimo especificado (s); e 2) Gerar regras usando conjuntos de itens frequentes (*itemsets*) que têm confiança maior ou igual à confiança mínima especificada (c).

A aplicação dessa técnica em bases de dados relacionadas a licitações tem a finalidade de identificar grupos de empresas associadas mediante a atuação frequente e conjunta entre elas. Segundo Silva (2011) [8], a evidência de grupos de empresas que participam frequentemente dos mesmos processos licitatórios é bastante relevante para detecção de carteis, pois demonstra uma determinada associação e cooperação entre elas, à luz da Teoria dos Jogos Repetidos. Tal cooperação se opõe ao princípio da igualdade em um procedimento licitatório e torna a competitividade entre os licitantes prejudicada, resultando em um possível aumento dos valores pagos pela Administração Pública. Uma regra de associação entre empresas pode ser descrita pela implicação (1), interpretada da seguinte maneira: se a empresa A participar de uma licitação, então há uma boa chance de participação da empresa B na mesma licitação. Logo, considera-se que a empresa A e a empresa B formam um grupo.

$$Empresa\,A \Rightarrow Empresa\,B \tag{1}$$

Existem diversos algoritmos na literatura para descoberta de regras de associação. Neste estudo, optou-se por utilizar o algoritmo *Apriori*, implementado no pacote *arules* do *R*. Para utilizar o algoritmo, foi necessário configurar os parâmetros de suporte e confiança mínimos, interpretados da seguinte maneira:

1. **Suporte (s)**: corresponde à probabilidade de a regra se repetir no conjunto de dados e, em geral, assume valor baixo para que boas regras não sejam descartadas. É calculado pelo quociente entre o número de registros com a participação das empresas A e B e o número total de registros ($s = \frac{n_{A,B}}{n}$). Segundo Silva (2011) [8], é razoável assumir que um cartel tenha atuado de 10 a 15 vezes num ciclo político de quatro anos. Neste estudo, foram testados valores de suporte considerando uma participação conjunta mínima de 5, 10 e 15 vezes. É necessário frisar que, apesar de

a chance de detecção de conluio aumentar ao se considerar regras com suportes altos, é possível que boas regras sejam ignoradas pelo algoritmo.

2. **Confiança (c)**: é uma medida que auxilia a identificar a qualidade de uma regra de associação (valores altos indicam boas regras). É calculada pelo quociente entre o número de registros com a participação das empresas A e B e o número de registros com a participação da empresa A ($c = \frac{n_{A,B}}{n_A}$). Neste estudo, considerou-se $c = 80\%$, percentual mais utilizado na literatura pesquisada.

Ademais, foi necessário preparar um *dataset* de entrada para o algoritmo *Apriori*. Para tanto, foram realizadas junções entre os *datasets* de licitações, contratos e participantes, bem como selecionadas apenas licitações que tiveram entre 2 e 10 participantes. As licitações que tiveram apenas 1 participante foram descartadas por não terem relevância nesta análise, uma vez que não houve possibilidade de concorrência no certame. Acerca do limite superior, decidiu-se por fazê-lo a fim de eliminar processos licitatórios que caracterizaram apenas um ambiente de concorrência alta, com chance baixa de colusão entre os participantes, conforme Fraga (2017) [12]. A partir do aludido *dataset*, foi criada uma matriz esparsa de tal forma que as linhas representassem os processos licitatórios e as colunas as empresas. Cada empresa assumiu o valor 1, se participou da licitação, ou vazio, caso contrário. Em seguida, essa matriz foi transformada em um objeto de transações (12.567 licitações) e itens (8.973 empresas), que serviu de entrada para o algoritmo.

Na etapa de treinamento, buscou-se por regras de associação considerando não só o estado do Paraná, mas também cada uma das partições encontradas pelo método de clusterização. Essa estratégia foi estabelecida com vistas a minimizar a influência do espaço de soluções sobre o cálculo de estatísticas de validação das regras. Presumiu-se que regras de baixa qualidade, identificadas ao se considerar o estado como um todo, passariam a apresentar uma qualidade melhor para discriminar indícios de cartel, quando detectadas a partir de partições do estado.

Na sequência, foram feitas algumas podas no conjunto de regras descoberto. Primeiramente, regras redundantes, por serem subconjuntos de outras regras cuja confiança foi a mesma ou maior, foram excluídas. Além disso, percebeu-se a existência de algumas regras duplicadas ao não se considerar a ordem entre empresas subsequentes e consequentes, como por exemplo: $A \Rightarrow B$ é igual a $B \Rightarrow A$. Levando em consideração que o objetivo desta análise foi tão somente formar grupos, sem a intenção de compreender a implicação decorrente, apenas uma dessas regras foi considerada. Por fim, foram excluídas regras compostas por pelo menos uma empresa cuja quantidade de participação foi maior ou igual ao percentil 99% da distribuição de participação de empresas no ciclo político de 2013 a 2016, buscando evitar padrões relacionados a grandes fornecedores que participaram coincidentemente das mesmas licitações.

2.4.2. Validação das Regras de Associação

É importante destacar que as regras de associação geradas devem ser avaliadas previamente à conclusão de que as empresas ali presentes sejam integrantes de um cartel, pois, muitas vezes, a característica de atuação frequente descoberta é apenas uma coincidência ou então vem do fato de que as empresas são de grande porte e participam com mais frequência de licitações. A seleção a partir dos parâmetros suporte e confiança do algoritmo *Apriori*, considerando o contexto de descoberta de carteis, não é, portanto, suficiente. Nesse sentido, as regras geradas passaram por uma etapa de refinamento a fim de classificá-las segundo o risco associado à existência de cartel.

Cada uma das regras selecionadas via suporte e confiança foi avaliada por meio de duas estatísticas de validação de sucesso contratual, quais sejam: 1) a probabilidade de o grupo de empresas pertencente à regra r firmar um contrato com a Administração Pública (Val_{1r}): quanto maior a probabilidade, maior é a chance de o grupo formar um cartel; e 2) a mediana das razões entre os valores dos contratos celebrados pelo grupo de empresas da regra r e os valores orçados pela Administração Pública (Val_{2r}): segundo Ishii (2009) [4], quanto mais próxima de 1 estiver essa razão, principalmente acima do limiar de 0,95, maior é a chance de o grupo pertencente à regra r participar de um jogo cooperativo entre empresas.

Na equação (2), $venc_r$ reflete o número de licitações vencidas por pelo menos uma das empresas da regra r, enquanto $disp_r$ representa o número de licitações disputadas pelo grupo de empresas pertencente à regra r. Em ambas as quantidades, todos os membros da regra r deveriam estar presentes na licitação para que a contagem fosse computada. Além disso, no caso de uma licitação loteada, foi computada uma unidade na quantidade $disp_r$, independentemente do número

de lotes, e uma unidade em $venc_r$ quando o grupo de empresas venceu pelo menos um dos lotes disputados.

$$Val_{1r} = \frac{venc_r}{disp_r} \quad (2)$$

A equação (3) se refere à mediana das razões entre os valores dos contratos ($vlvenc_{ir}$) e os valores orçados ($vlorçado_{ir}$), considerando somente licitações vencidas por pelo menos uma das empresas pertencentes à regra r ($venc_r$), quando todo o grupo participou. Na literatura, não foram encontrados estudos em que essa medida foi estabelecida como critério de validação de regras de associação. Contudo, a relação entre os valores contratado e orçado se mostrou bastante relevante para identificação de cartel em estudos anteriores, o que corroborou para seu uso como uma medida de avaliação e seleção de regras.

$$Val_{2r} = \underset{i=1,2,\ldots,venc_r}{\text{mediana}} \left\{ \frac{vlvenc_{ir}}{vlorçado_{ir}} \right\} \quad (3)$$

Por fim, no intuito de melhor caracterizar o risco de cartel associado a cada regra por meio de um único escore, foi construído um indicador contínuo de risco baseado na medida *Weighted F-measure*, comumente utilizada para avaliar a qualidade de classificadores. Considerando que: 1) Essa medida busca um equilíbrio entre duas estatísticas por meio do cálculo de uma média harmônica; e 2) O risco associado às duas estatísticas sob estudo (Val_{1r} e Val_{2r}) cresce à medida que se aproximam de 1, o uso da medida *Weighted F-measure*, no contexto deste trabalho, se mostrou bastante relevante em termos de sentido prático. Ou seja, quanto maior o valor da medida *Weighted F-measure*, maior é o risco de cartel associado à regra r. A referida medida foi definida pela seguinte equação 4:

$$Risco_r = F_{\beta r} = \frac{(1+\beta^2).Val_{1r}.Val_{2r}}{Val_{1r} + \beta^2.Val_{2r}} \quad (4)$$

$0 < \beta < 1$ dá mais peso para a estatística Val_{2r}, ao passo que $\beta > 1$ dá mais peso para a estatística Val_{1r}. Neste trabalho, definiu-se β = 0,5, ou seja, a estatística Val_{2r} teve duas vezes mais peso que a estatística Val_{1r} na construção do escore de risco. Na literatura, não foram encontrados estudos em que houve uma comparação direta entre as duas estatísticas, porém, assumiu-se que o sucesso das empresas de um cartel baseado no montante lucrado, mesmo tendo contratado poucas vezes, seria mais importante que o sucesso de um cartel baseado na alta proporção de vezes que contratou, desconsiderando um possível lucro baixo percebido.

2.5. Fatores de Risco de Cartel

A partir das bases de dados disponíveis, foi estabelecido um conjunto de possíveis fatores representativos da prática de atividades ilícitas em licitações. O levantamento foi realizado por meio de consulta a pesquisas científicas na literatura e a órgãos governamentais responsáveis pela disseminação de boas práticas no combate a fraudes (Tabela 1): Ministério da Justiça (2008) [3], OCDE (2009) [5], Morais (2016) [11], Sales (2016) [25], Tóth (2015) [26] e Centro Internacional de Recursos Anti-Corrupção (IACRC) (2019) [27].

Tabela 1: Fatores de Risco

Código	Descrição
propValorPago	Proporção do valor contratado em relação ao orçado
taxaEmprVenc	Taxa de contratos firmados pela empresa vencedora
taxaMunicEmprVenc	Taxa de sucesso municipal da empresa vencedora
quantPartic	Quantidade de participantes
partEmprVenc	Número de participações da empresa vencedora
idadeEmprVenc	Data de constituição da empresa vencedora
ativEmprVenc	Quantidade de atividades da empresa vencedora
taxaVencOutras	Taxa de contratos firmados pela empresa vencedora em relação as demais participantes
propDesclassificadas	Proporção de empresas desclassificadas ou inabilitadas
taxaEmprNunca	Parcela de empresas participantes que nunca venceram contratos
socios	Sócios em comum
mesmoTel	Empresas com mesmo número de telefone

Fonte: O Autor

2.6. *Self-Organizing Maps (SOM)*

O algoritmo *Self-Organizing Maps* (SOM) é uma rede neural artificial sem camadas escondidas em que o aprendizado é não supervisionado, funcionando como um mapeamento direto entre o conjunto de treinamento e a rede de saída. É normalmente utilizado para representar um número grande de dimensões em um espaço bi ou tri-dimensional, sendo possível observar associações e correlações entre atributos e realizar uma análise mais apurada do montante de dimensões, sempre mantendo a originalidade dos dados. Além do mais, a técnica SOM faz com que registros similares fiquem próximos uns aos outros na representação dimensional, permitindo criar agrupamentos por meio da identificação de regiões distintas.

Kind (2013) [28] explica que um dos aspectos mais importantes desse algoritmo é o fato de conseguir aprender a classificar sem a necessidade de um atributo de exemplo, característica inerente a classificadores supervisionados. Além disso, outra característica importante de um SOM é que a fase de treinamento é um processo competitivo, chamado quantização vetorial, em que cada nó ou neurônio no mapa compete com outros a fim de se tornar mais semelhante aos dados de treinamento, ou seja, cada neurônio tenta retratar o melhor possível o conjunto de dados de treinamento. Essa última característica e o fato de levar em consideração uma função de vizinhança entre os neurônios faz com que o SOM mantenha a configuração multidimensional dos dados.

Na Figura 1, é possível observar uma representação esquemática de como um SOM é treinado, em que um conjunto de dados de treino de tamanho n é representado em um espaço bi-dimensional de k neurônios, acompanhado por um vetor de pesos (w) de tamanho m igual ao número de atributos que caracterizam cada registro: $\vec{w}_j = [w_{j1}, w_{j2}, ..., w_{jm}]$. Considerando os dados deste trabalho, cada uma das n licitações do conjunto de treinamento foi considerada uma parte da camada de entrada, sendo acompanhada por um vetor de características de tamanho m (12 variáveis): $\vec{lic}_i = [fator_{i1}, fator_{i2}, ..., fator_{i12}]$.

Cada um dos neurônios do mapa bi-dimensional é composto por um vetor de pesos de mesma dimensão do número de características e pode refletir mais de um dos n registros. No processo de treinamento, os registros do conjunto de treinamento são individualmente utilizados para corrigir os vetores de peso, de modo que o neurônio que melhor representa o registro

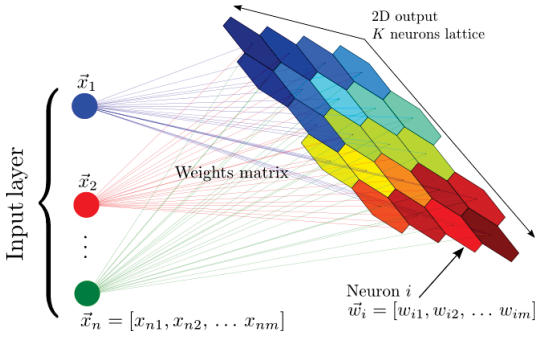

Figura 1: Representação esquemática de um SOM (KIND, 2013)

seja modificado, juntamente com os vetores de pesos de neurônios vizinhos, sendo tal processo repetido para todos os registros.

Para aplicação do aludido algoritmo nos dados sob estudo, foi utilizada a biblioteca *kohonen* do *software R*, que dispõe de funções para treinamento e visualização gráfica de resultados. O SOM foi treinado considerando uma rede neural de tamanho 15 x 15 (225 neurônios), um formato hexagonal para os neurônios, vizinhança gaussiana e representação do mapa em um plano toroidal. No tocante aos dados, considerou-se como objeto de entrada uma matriz de tamanho 1.173 licitações com suspeitas de atuação de cartel (somente registros sem *missings* nos indicadores de cartel, tanto do Paraná quanto dos mercados de licitação) *versus* 12 atributos, padronizados via estatística *min-max*, dada por $\tilde{x} = \frac{x - min(x)}{max(x) - min(x)}$. Todo o processo de treinamento foi realizado 3.200 vezes a fim de otimizar a configuração do mapa final.

3. Resultados e Discussões

3.1. Identificação de Mercados de Licitação

Na Figura 2, é possível observar os fluxos origem-destino das 658 empresas que participaram de 20 ou mais processos licitatórios em relação aos 11 gepatrias do Paraná. Linhas mais espessas com cores mais fortes indicam uma quantidade maior de participação. As empresas estão representadas pelo ícone de uma casa e simbolizam a origem do fluxo. Já os 11 gepatrias estão denotados pelos polígonos em cinza, sendo o centroide de cada um o destino do fluxo.

Com base nos fluxos, nota-se que algumas empresas tiveram uma abrangência estadual de participação, principalmente as situadas no gepatria de Curitiba e com fluxos em direção a diversas localidades do estado, não mantendo uma atuação apenas na respectiva re-

gião. Porém, percebe-se que as linhas são mais finas, salvo algumas exceções de empresas localizadas na capital paranaense.

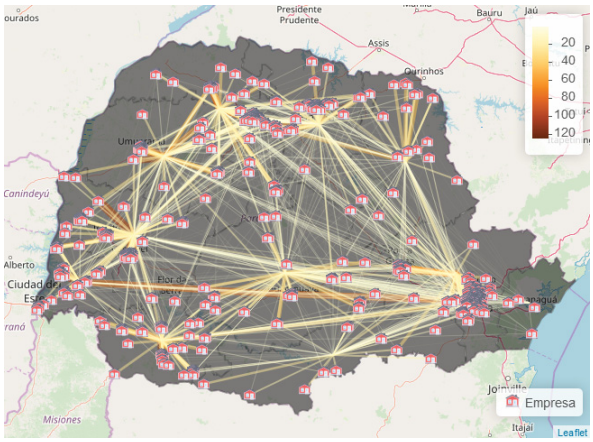

Figura 2: Mapa de fluxo origem-destino das 658 empresas com mais de 20 participações em relação aos 11 gepatrias do Paraná

Em contrapartida, observa-se que muitas empresas participaram de licitação de forma regionalizada, não se afastando sobremaneira de sua sede, característica mais recorrente no oeste do estado, onde se encontram municípios como Cascavel e Foz do Iguaçu. Isso pode ser evidenciado por meio das linhas mais espessas com cores fortes no perímetro do entorno de cada gepatria, indicando uma concentração maior de participação. Logo, é válido pensar em uma divisão territorial dos gepatrias a partir das participações das empresas, vislumbrando a formação de possíveis mercados de licitação.

A partir da execução dos algoritmos de clusterização, concluiu-se que os métodos SKATER e hierárquicos não apresentaram resultados satisfatórios e de fácil interpretação, sendo desconsiderados do processo. Já o método k-means forneceu resultados bastante interessantes do ponto de vista prático e ministerial, considerando uma padronização dos dados tanto por linhas quanto pela estatística *z-escore*. O algoritmo aglomerou as empresas em três grupos distintos, cujas médias de participação em cada gepatria, tanto absoluta quanto percentual, podem ser visualizadas na Tabela 2. Em apenas duas das 11 regiões, Guarapuava e União da Vitória, não foi possível encontrar, em ambas as formas de normalização, diferenças estatísticas em relação às médias de participação por grupo, considerando um nível de significância de 5%.

Além disso, detectou-se uma divergência na classificação do gepatria de Guarapuava. A partir da padronização *z-escore*, Guarapuava foi classificado no grupo 3 (G3), ao passo que pela padronização por linhas, houve uma alteração para o grupo 2 (G2). Essa mudança pode ter sido decorrente da localização centralizada do gepatria, que possibilitou às empresas situadas nas demais regiões do entorno um acesso mais fácil para participar de procedimentos licitatórios. Contudo, segundo especialistas do MPPR, a região de Cascavel tem mais abrangência sobre Guarapuava do que a região de Curitiba, sendo o resultado advindo da padronização por linhas mais representativo.

A associação de cada gepatria a um dos três grupos foi realizada por meio da seleção da maior média dos percentuais de participação das empresas, apresentadas na coluna Padronização B da Tabela 2. A distribuição territorial encontrada pode ser visualizada na Figura 3. É necessário ressaltar que, mesmo não tendo considerado uma continuidade espacial, os grupos encontrados foram formados por gepatrias vizinhos geograficamente, o que já era esperado considerando a interpretação do mapa de fluxo origem-destino presente na Figura 2. Nas Figuras 4, 5 e 6, é possível observar os fluxos origem-destino das empresas participantes de licitação nos grupos 1, 2 e 3, respectivamente. Nota-se que empresas mais distantes dos respectivos grupos de atuação participaram de poucos processos licitatórios, enquanto que as mais próximas, inclusive localizadas na região, tiveram uma quantidade de participação mais elevada, o que valida a configuração de grupos encontrada.

3.2. Detecção de Potenciais Carteis

Na Tabela 3, são apresentados os resultados relativos ao processo de descoberta de indícios de cartel por meio do treinamento de regras de associação, considerando o estado do Paraná e os três mercados de licitação. Foram testados três valores de suporte para o estado como um todo e apenas o suporte com no mínimo cinco participações em licitações para análise dos *clusters*. No que se refere ao estado do Paraná, à medida que se aumentou o valor de suporte, o número de regras encontradas e selecionadas diminuiu, porém as médias das estatísticas de avaliação e do risco de cartel aumentaram consideravelmente.

Ademais, ainda no âmbito do estado e considerando o suporte de $\frac{5}{12567}$ e os critérios de poda configurados, 524 regras encontradas pelo algoritmo *Apriori* foram eliminadas do processo, perfazendo um percentual de 32% de regras selecionadas. Já os suportes de $\frac{10}{12567}$ e $\frac{15}{12567}$ apresentaram percentual de regras selecionadas de 34% e 33%, respectivamente. A confiança média se

Figura 3: Mercados de Licitação identificados via clusterização de empresas utilizando o algoritmo k-means com padronização dos dados por linhas

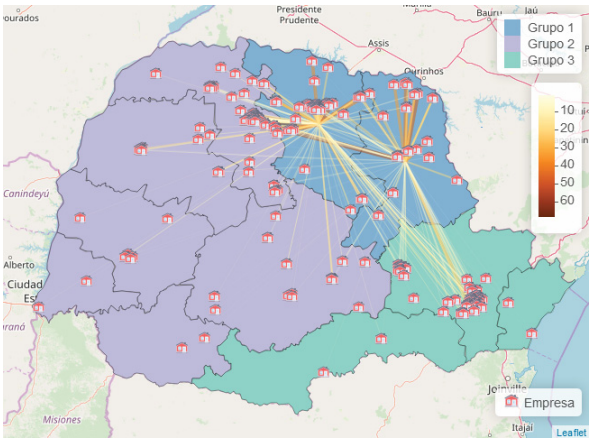

Figura 4: Mapa de fluxo origem-destino das empresas participantes de licitação no grupo 1, composto pelos gepatrias de Londrina e Santo Antônio da Platina

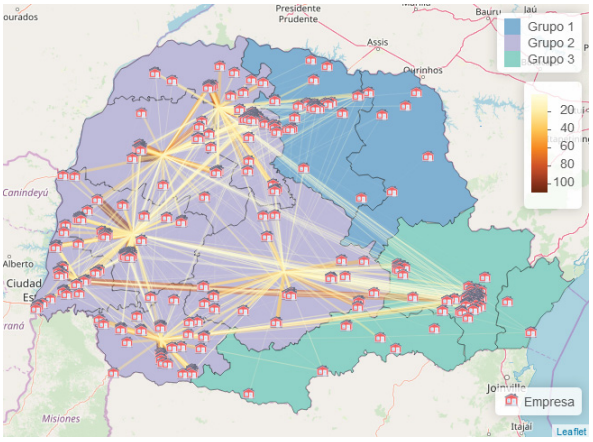

Figura 5: Mapa de fluxo origem-destino das empresas participantes de licitação no grupo 2, composto pelos gepatrias de Cascavel, Francisco Beltrão, Foz do Iguaçu, Guarapuava, Maringá e Umuarama

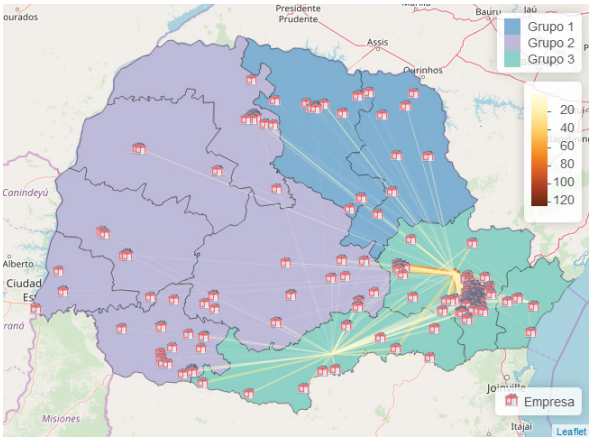

Figura 6: Mapa de fluxo origem-destino das empresas participantes de licitação no grupo 3, composto pelos gepatrias de Curitiba, Paranaguá e União da Vitória

manteve no patamar de 93%, em todos os níveis de suporte, indicando uma boa qualidade das regras no que se refere à execução do algoritmo.

Dentre os três mercados de licitação, o G2, além de apresentar o maior risco médio de cartel, foi o grupo que obteve mais regras selecionadas, mesmo tendo uma quantidade total de regras geradas pelo algoritmo menor que o G3. As regras do G3, em sua maior parte, foram excluídas pelo fato de serem compostas por grandes fornecedores que participaram em mais de 44 licitações, percentil 99% da distribuição de participação de empresas no ciclo político de 2013 a 2016. Ou seja, muitas das empresas que participaram de procedimentos licitatórios no G3, principalmente no gepatria de Curitiba, foram enquadradas como grandes fornecedores, cuja presença nas regras seria devido a uma coincidência, ao acaso ou ao fato de serem provedoras exclusivas do serviço ou produto contratado. Além disso, nota-se uma discrepância em relação aos percentuais de regras selecionadas entre os três mercados. Os grupos G1 e G2 tiveram um percentual de 40% e 48% de regras selecionadas, respectivamente, ao passo que o G3 teve apenas 19%, característica também relacionada à participação de grandes fornecedores no gepatria de Curitiba.

A partir das regras selecionadas, associou-se um escore de risco a cada uma das 25.169 licitações (34.512 contratos), sendo possível ordená-las em uma matriz de risco de cartel. No que concerne ao Paraná, 1.364 licitações (3.631 contratos) tiveram pelo menos uma das 245 regras aplicadas, perfazendo um percentual de 5,4%. Em relação aos mercados de licitação, foram

Tabela 2: Média das quantidades absolutas e percentuais de participação dos três grupos de empresas por gepatria, considerando os resultados obtidos pelo algoritmo k-means e uma padronização tanto pela estatística *z-escore* (A) quanto por linhas (B)

Gepatria	Padronização A			Padronização B		
	G1	G2	G3	G1	G2	G3
Cascavel +	1,9	8,2	0,7	0,001	0,2	0,007
Curitiba +	2,2	0,5	25	0,04	0,02	0,82
Foz do Iguaçu +	0,2	3,1	0,1	0	0,07	0,001
Francisco Beltrão +	0,4	4,9	0,6	0	0,13	0,002
Guarapuava -	1,1	2,2	3,4	0,001	0,07	0,02
Londrina +	11,4	4,1	1,1	0,15	0,11	0,01
Maringá +	5,3	7,2	0,6	0,006	0,18	0,004
Paranaguá +	0,1	0,05	1,5	0,001	0,007	0,04
Santo Antô. Platina +	25,9	0,7	1	0,79	0,02	0,01
Umuarama +	2,1	5,1	0,3	0	0,11	0,002
União da Vitória *	0,2	0,4	4,2	0,001	0,04	0,05

Notas: Padronização A: k-means com padronização dos dados pela medida *z-escore* (média das quantidades absolutas); Padronização B: k-means com padronização dos dados por linhas (média das quantidades percentuais);
+ significativo a 0,05 tanto na Padronização A quanto na Padronização B;
* significativo a 0,05 somente na Padronização A;
- significativo a 0,05 somente na Padronização B.

Tabela 3: Quadro resumo da aplicação do algoritmo *Apriori* considerando o Estado do Paraná e os três *clusters* de gepatrias identificados via k-means

	Paraná (suportes)			G1	G2	G3
	$\frac{5}{12567}$	$\frac{10}{12567}$	$\frac{15}{12567}$	$\frac{5}{2169}$	$\frac{5}{7947}$	$\frac{5}{2451}$
Total de regras	769	64	21	83	310	403
Selecionadas	245	22	7	33	149	76
Confiança média	0,93	0,93	0,92	0,93	0,94	0,93
Val_1 médio	0,72	0,88	0,91	0,70	0,81	0,57
Val_2 médio	0,83	0,87	0,92	0,84	0,85	0,81
Risco médio	0,79	0,86	0,91	0,77	0,83	0,72

Notas: G1 = Londrina e Santo Antônio da Platina; G2 = Cascavel, Foz do Iguaçu, Francisco Beltrão, Guarapuava, Maringá e Umuarama; G3 = Curitiba, Paranaguá e União da Vitória.

detectadas 230 licitações (5,3%) com indícios de cartel no grupo G1, 814 (4,9%) no G2 e 279 (6,6%) no G3. Além disso, 191 licitações suspeitas selecionadas por meio da análise macro do estado do Paraná não foram identificadas via análise dos mercados de licitação. Em contrapartida, 150 processos licitatórios com indícios de conluio foram detectados somente a partir dos grupos de gepatrias. As contratações públicas oriundas de licitações suspeitas no Paraná custaram aos cofres públicos o equivalente a R$ 659.638.750,00 milhões, cerca de 7,5% do total contratado pelos municípios no período. Ainda, considerando apenas as licitações suspeitas identificadas a partir dos mercados de licitação, o custo passou a ser de R$ 658.521.696,00 milhões, pouco mais de um milhão de diferença em relação ao total do estado.

Na Figura 7, é possível observar as distribuições de densidade relacionadas ao indicador de risco de cartel, tanto para o estado quanto para os três grupos. Além de as densidades serem parecidas e apresentarem assimetria à esquerda, percebe-se uma grande massa de dados concentrada em um patamar acima do limiar de 0,7 de risco, indicando fortes evidências progressivas de conluio na maioria das licitações selecionadas via regras de associação, principalmente aquelas cujo risco associado se encontra acima de 0,9 (365 licitações no Paraná, 56 no G1, 274 no G2 e 88 no G3). A média de risco de cartel considerando o estado como um todo foi de 0,838. Nos mercados de licitação G1, G2 e G3, esse valor passou a ser de 0,83, 0,839 e 0,842, respectivamente. Isso demonstra que, muito embora 341 licitações supostamente cartelizadas foram descobertas em apenas um dos recortes espaciais, 191 no estado e 150 nos grupos, no geral, os resultados de ambas as configurações foram parecidos.

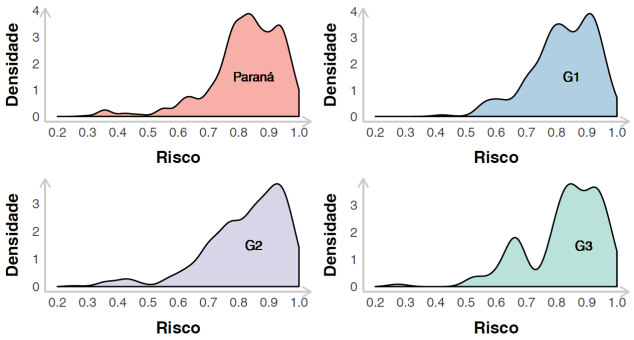

Figura 7: Distribuições de densidade do indicador de risco de cartel (F_β), considerando o Paraná e os três mercados de licitação. G1 = Londrina e Santo Antônio da Platina; G2 = Cascavel, Foz do Iguaçu, Francisco Beltrão, Guarapuava, Maringá e Umuarama; G3 = Curitiba, Paranaguá e União da Vitória

3.3. Relação com Fatores de Risco de Cartel

Os *heatmaps* da Figura 8 mostram as relações entre os dois indicadores de indícios de cartel e dez tipologias de risco. A graduação de cores, do amarelo ao azul, representa a variação de cada atributo, e a posição de cada neurônio é a mesma em todos os *heatmaps*. Nos *heatmaps* 1 e 2, é possível perceber uma associação entre os riscos de cartel identificados a partir do Estado e dos três mercados de licitação, ficando evidente a semelhança de resultados de ambos os espaços de

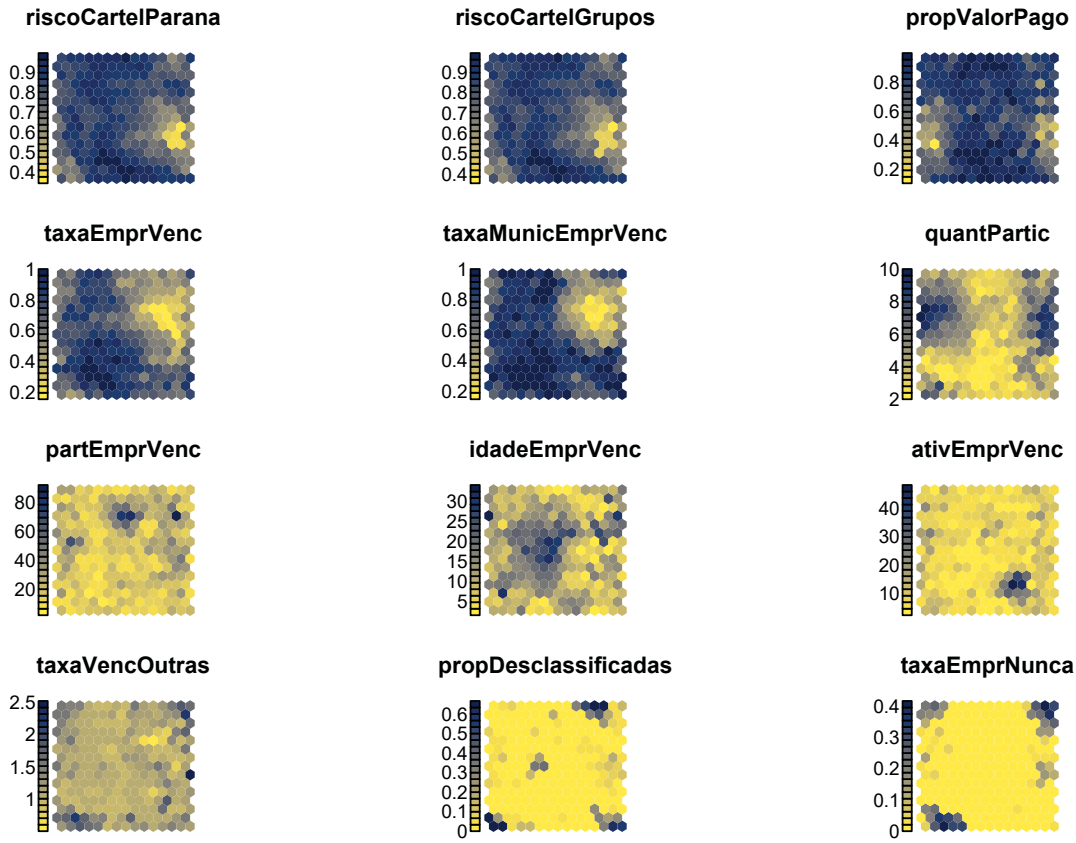

Figura 8: Relação de dez tipologias de risco com os indicadores de cartel. riscoCartelParana (1), riscoCartelGrupos (2), propValorPago (3), taxaEmprVenc (4), taxaMunicEmprVenc (5), quantPartic (6), partEmprVenc (7), idadeEmprVenc (8), ativEmprVenc(9), taxaVencOutras (10), propDesclassificadas (11) e taxaEmprNunca (12)

solução, o que também pôde ser evidenciado na Figura 7.

Além disso, as variáveis *taxaEmprVenc* e *taxaMunicEmprVenc*, correlacionadas entre si, apresentam uma tendência de associação positiva com os indicadores de risco de cartel. Ou seja, à medida que se aumenta a taxa de vitórias da empresa vencedora da licitação, quer seja no geral ou levando em consideração somente o município contratante, espera-se que os riscos de cartel também aumentem. Em contrapartida, ocorre uma tendência de associação negativa entre a quantidade de participantes na licitação (*quantPartic*) e os riscos de cartel, o que já era esperado considerando que ambientes com poucos competidores contribuem para uma atuação colusiva de um cartel. Ademais, a partir dos *heatmaps* 10, 11 e 12, percebe-se que a maioria dos neurônios apresentam características propícias para um ambiente de cartel: proporção de empresas desclassificadas (*propDesclassificadas*) e taxa de empresas que nunca venceram uma licitação (*taxaEmprNunca*), em sua grande parte, nulas; e razão entre a taxa de vitória da empresa vencedora em relação à taxa de vitória das outras empresas participantes (*taxaVencOutras*) próxima de 1.

4. Conclusão

Propôs-se, neste artigo, uma metodologia para detecção de indícios de cartel em licitações atinentes à área de engenharia, considerando a gestão municipal paranaense de 2013 a 2016. A partir de dados públicos e de técnicas estatísticas e de *machine learning*, foi possível identificar mercados de licitação, encontrar regras de associação mediante a atuação conjunta e frequente de empresas e criar um indicador de risco de cartel.

Os resultados advindos da confecção de mapas de fluxo origem-destino e da aplicação da técnica k-means possibilitaram uma melhor compreensão da atuação de empresas participantes de licitação nos municípios do Paraná. A partir dos mapas, observou-se que muitas empresas participaram de licitação de forma regionalizada, não se afastando sobremaneira de seu município de origem. Essa informação foi de grande valia para

este estudo, pois confirmou a necessidade de dividir o estado em espaços menores de solução. O uso da técnica k-means, por sua vez, proporcionou a identificação de grupos de empresas atuantes em três regiões do estado, que serviram como recortes espaciais para as posteriores análises.

A utilização do algoritmo *Apriori* para geração de regras de associação viabilizou a identificação de potenciais carteis com base no padrão associativo de participação frequente e conjunta de empresas. Assim como em estudos anteriores, foi necessário realizar uma série de procedimentos para minimizar a possibilidade de coincidências na caracterização das atividades colusivas. Para tanto, além de considerar métodos já praticados por outros autores, foram propostos novos critérios de poda, bem como uma nova estatística de validação de regras, baseada no cálculo da proporção do valor pago em relação ao valor orçado pela Administração Pública. Em estudos futuros, buscar-se-á a implementação de novas estatísticas de validação, como a inclusão da quantidade de concorrentes em licitações das quais empresas suspeitas de conluio participaram. As regras de associação identificadas nos três mercados de licitação foram bastante similares às encontradas no Paraná, demonstrando que o treinamento do algoritmo considerando somente o estado já seria suficiente para detecção de eventuais carteis. Ainda, a partir do treinamento da rede neural SOM, foram observados diferentes tipos de associação entre o risco de cartel e tipologias de irregularidades em processos licitatórios ou contratos. O conhecimento adquirido a partir dessa análise foi bastante interessante do ponto de vista técnico e prático, devido à sua fácil implementação e interpretabilidade visual.

Por fim, cabe salientar que o uso de métodos estatísticos e de *machine learning* se mostrou bastante promissor para a detecção de padrões de associação frequente entre empresas licitantes. A identificação de fornecedores cooperantes entre si, a partir de procedimentos criteriosos e científicos, respalda a possibilidade de uso do conhecimento adquirido no MPPR, principalmente no auxílio e assessoramento da atividade-fim ministerial, quando da tomada de decisão, propiciando uma atuação proativa e otimizada em relação à atuação colusiva de empresas. Entretanto, é necessário frisar que o aprendizado alcançado fornece tão somente indícios de suspeição da existência de carteis, podendo ser utilizado apenas para fins de inteligência e não para produção de provas em processos judiciais.

Agradecimentos

P.J.S.C.R. gostaria de agradecer ao Professor Walmes M. Zeviani pelo apoio e auxílio na resolução deste trabalho, à Coordenação da Pós-Graduação em Data Science & Big Data, à sua família e a todos os integrantes do Núcleo de Inteligência do Ministério Público do Estado do Paraná.

Referências

[1] Presidência da República, Casa Civil, Subchefia para Assuntos Jurídicos, *Lei nº 8.666, de 21 de junho de 1993*, Disponível em: Lei nº 8.666. Acesso em: mar/2019

[2] B. Carazza, *Interesses Econômicos, Representação Política e Produção Legislativa no Brasil sob a Ótica do Financiamento de Campanhas Eleitorais*, Tese de Doutorado, Faculdade de Direito da UFMG (2016)

[3] Ministério da Justiça, Secretaria de Direito Econômico, Departamento de Proteção e Defesa Econômica, *Combate a Carteis em Licitações - Guia prático para pregoeiros e membros de comissões de licitação*, (Coleção SDE/DPDE 02/2008)

[4] R. Ishii, *Favor exchange in collusion: Empirical study of repeated procurement auctions in Japan*, International Journal of Industrial Organization, pag. 137-144 (2009)

[5] Organização para a Cooperação e Desenvolvimento Econômico, *Diretrizes para Combater o Conluio entre Concorrentes em Contratações Públicas*, Fev/2009

[6] Presidência da República, Casa Civil, Subchefia para Assuntos Jurídicos, *Lei nº 8.884, de 11 de junho de 1994*, disponível em: Lei nº 8.884, acesso em: mar/2019

[7] Presidência da República, Casa Civil, Subchefia para Assuntos Jurídicos, *Lei nº 8.137 de 27 de dezembro de 1990*, disponível em: Lei nº 8.137, acesso em: mar/2019

[8] C. V. S. Silva, C. G. Ralha, *Agentes de Mineração e sua Aplicação no Domínio de Auditoria Governamental*, Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Ciência da Computação (2011)

[9] S. S. Padhi, P. K. K. Mohapatra, *Detection of collusion in government procurement auctions*, Journal of Purchasing & Supply Management, pag. 207-221 (2011)

[10] I. Morozov, E. Podkolzina, *Collusion Detection in Procurement Auctions*, National Research University, Higher School of Economics (2013)

[11] C. M. M. Morais, *Proposição de indicadores para investigação de licitações por meio de técnicas de reconhecimento de padrões estatísticos e mineração de dados*, Universidade de Brasília, Faculdade de Tecnologia, Departamento de Engenharia Elétrica (2016)

[12] A. A. Fraga, H. M. B. Ramalho, A. T. C. Almeida, *Detecção de casos suspeitos de fraudes em licitações realizadas nos municípios da Paraíba: uma aplicação de técnicas de mineração de dados*, Universidade Federal da Paraíba, Centro de Ciências Sociais Aplicadas,

Programa de Pós-Graduação em Economia do Setor Público (2017)
[13] Tribunal de Contas do Estado do Paraná, *Lei Complementar nº 113 de 15/12/2015*, disponível em: Lei Complementar nº 113, acesso em: mar/2019
[14] Tribunal de Contas do Estado do Paraná, *Portal de Informações para Todos*, disponível em: Portal de Informações para Todos, acesso em: mar/2019
[15] R Core Team (2016), *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), disponível em: R-project, acesso em: mar/2019
[16] P. J. S. C. Rosa, *Biblioteca LicitaR*, disponível em: LicitaR, acesso em: mar/2019
[17] *API Receita WS*, disponível em: Receita WS, acesso em: mar/2019
[18] Controladoria-Geral da União, *Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS)*, disponível em: CEIS, acesso em: mar/2019
[19] Tribunal Superior Eleitoral, *Repositório de Dados Eleitorais*, disponível em: Repositório de Dados Eleitorais, acesso em: mar/2019
[20] Instituto Brasileiro de Geografia e Estatística, *Bases e Referenciais, Bases Cartográficas e Malhas Digitais*, disponível em: Malhas Digitais, acesso em: mar/2019
[21] R. M. Assuncao, M. C. Neves, G. Camara, C. da C. Freitas, *Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees*, International Journal of Geographical Information Science, Vol. 20, Nº. 7, Agosto de 2006, pag. 797-811
[22] L. P. Fávero, P. Belfiore, F. L. da Silva, B. L. Chan, *Análise de Dados: Modelagem Multivariada para Tomada de Decisões*, Rio de Janeiro: Elsevier (2009)
[23] R. Goldschmidt, E. Passos, E. Bezerra, *Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações*, 2ª Edição, Rio de Janeiro: Elsevier (2015)
[24] R. Agrawal, R. Srikant, *Fast Algorithms for Mining Association Rules*, Proceedings of the 20th VLDB Conference, Santiago, Chile (1994)
[25] L. J. Sales, R. T. de Menezes, *Proposta de modelo de classificação de risco de contratos públicos*, Universidade Nacional de Brasília, Faculdade de Economia, Administração e Contabilidade, Programa de Pós-Graduação em Economia (2016)
[26] B. Tóth, M. Fazekas, A. Czibik, I. J. Tóth, *Toolkit for detecting collusive bidding in public procurement: with examples from Hungary*, Corruption Research Center Budapest (2015)
[27] Centro Internacional de Recursos Anti-Corrupção (IACRC), *Guide to Combating Corruption & Fraud in Development Projects - Collusive Bidding Schemes* (2019)
[28] M. C. Kind, R. J. Brunner, *SOMz: photometric redshift PDFs with self organizing maps and random atlas*, Department of Astronomy, University of Illinois, Urbana, USA (2013)