Multicriteria Anomaly Detection in Government Purchases

Patrícia Maia[1,2], Wagner Meira Jr.[1], Breno Barbosa[2], Gustavo Cruz[2]

[1] Universidade Federal de Minas Gerais, Brazil
patricia.maia, meira, @dcc.ufmg.br
[2] Controladoria Geral da União, Brazil
patricia.maia,breno.barbosa, gustavo.cruz,@cgu.gov.br

Abstract. Government purchases are the usual instrument for public acquisition of goods and services. Despite extensive legislation and several control and auditing mechanisms, frauds are still diverse and commonplace at all levels of public administration. This work proposes a methodology for detecting anomalies in government purchases. The methodology promotes several levels of filtering with respect to entities involved and purchases considered as fraudulent considering diverse criteria. The applicability and effectiveness of the methodology is demonstrated through a case study using real data where we were able to identify a long term provider collusion.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: anomaly detection, government purchases

1. INTRODUÇÃO

No trabalho de auditoria acerca de procedimentos licitatórios, um dos principais objetivos é a verificação da efetiva ocorrência de competição entre os licitantes, proporcionando à Administração Pública a contratação em condições mais vantajosas no que diz respeito ao valor pago por bens e serviços, ou o valor recebido no caso de alienações. No caso da modalidade Pregão, instituída pela Lei 10.520/2002, tal verificação é facilitada pela possibilidade de serem observados os lances formulados pelas empresas licitantes na plataforma do Sistema Integrado de Administração de Serviços Gerais – SIASG, nos moldes de um leilão, com encerramento aleatório determinado pelo sistema.

Existem alguns indicativos multicritério de que o caráter competitivo da licitação pode ter sido frustrado por meio da limitação artificial da concorrência e que, sem uma análise de um volume expressivo de dados, seria difícil a caracterização de um padrão de comportamento anti-competitivo previamente combinado entre as empresas, resultando em divisão de mercado, que é um dos objetivos da cartelização.

A maioria dos indicativos dificilmente pode ser percebido na análise individual de procedimentos licitatórios, devido a vários fatores, dentre os quais destacam-se: a) A equipe de auditoria que realiza o trabalho em um ano pode ser diferente da equipe do ano seguinte, o que favorece a perda da memória dos trabalhos anteriores; b) O escopo da auditoria pode variar de um ano para outro, não havendo necessariamente a análise de procedimentos licitatórios envolvendo o mesmo objeto; c) A contratação de serviços continuados (como limpeza e vigilância) pode ser prorrogada por até 60 meses, o que faz com que essas contratações fiquem um período razoável sem serem auditadas; d) O procedimento licitatório analisado individualmente pode ter apenas um lote em disputa, não permitindo a análise da divisão da licitação entre os licitantes; e) São escaladas diversas equipes de auditoria para trabalhos

em vários órgãos públicos, não havendo necessariamente comunicação entre elas posteriormente à realização dos trabalhos para avaliação de possíveis práticas anti-competitivas em licitações; f) Não há o registro, por parte das equipes de auditoria, das ocorrências verificadas em análise de licitações em um banco de dados específico; g) Uma equipe que realiza trabalho de auditoria em um determinado órgão público pode ser escalada para um trabalho em outro órgão onde uma determinada empresa não atue, tornando mais difícil a identificação da divisão de mercado entre empresas; h) Embora seja desejável, não há homogeneidade quanto ao nível de experiência e/ou conhecimento técnico das equipes de auditoria.

Portanto, embora alguns indicativos possam ser verificados em um procedimento licitatório analisado individualmente, como, por exemplo, empresas que celebram contrato sem terem sido as que ofereceram melhores preços, a verificação da existência de um padrão a longo prazo é um desafio pelos fatores detalhados acima. Mais ainda, outros indicativos não são possíveis de serem identificados em trabalhos isolados de auditoria, como a permanência de empresas ao longo do tempo em determinados órgãos e o percentual de mercado de cada empresa. Assim, a análise de grandes volumes de dados surge como uma importante estratégia no trabalho de auditoria por permitir a identificação de padrões multicritério ao longo do tempo, os quais estão fortemente associados a comportamentos anti-competitivos por parte dos licitantes, direcionando os trabalhos de auditoria para aqueles procedimentos licitatórios com maior probabilidade de ocorrência de fraude.

Em suma, este trabalho propõe uma metodologia para detecção de anomalias, buscando possíveis conluios entre empresas licitantes. Demonstramos a aplicabilidade e a efetividade da metodologia proposta analisando pregões eletrônicos para fornecimentos de bens ou serviços à órgãos públicos situados no Estado de Minas Gerais. Através da utilização de mineração de padrões frequentes e correlação de series temporais multivariadas e análise conjugada de multicritérios nos processos de licitação, na modalidade pregão, foram reveladas situações que sugerem a ocorrência de fraude ou frustação ao caráter competitivo do certame, possivelmente por meio de ajuste ou combinação entre os concorrentes.

2. TRABALHOS RELACIONADOS

Diversas técnicas de detecção de anomalias vem sendo aplicadas considerando o contexto temporal [Gutflaish et al. 2017] [Tian et al. 2019] [Siddiqui et al. 2018] [Song et al. 2018]. Hallac, [Hallac et al. 2017] por exemplo, utilizou agrupamento e segmentação de séries temporais multivariadas para analisar diferentes características de um mesmo objeto buscando encontrar padrões nos dados que representem comportamentos particulares desse objeto. Yagoubi [YAGOUBI et al. 2018] propõe um *Multi-Scale Convolutional Recurrent Encoder-Decoder (MSCRED)* capaz de detectar anomalias e diagnosticar a causa raiz em series temporais multivariadas. O trabalho proposto por [Siddiqui et al. 2018] incorporou o feedback do analista para reduzir os falsos positivos e identificar quais as principais anomalias que devem ser investigadas. Esse feedback é utilizado para ajustar o ranking de anomalias depois de cada interação com o usuário, movendo para o topo do ranking aquelas anomalias de maior interesse pelo usuário. Vários estudos têm atuado no campo de detecção de conluio e fraudes em procedimentos licitatórios, dentre eles, Ralha e Sarmento Silva [Ghedini Ralha and Sarmento Silva 2012] analisaram as licitações do Governo Federal através de mineração de padrões frequentes e sistemas multiagentes e observaram a formação de cartéis envolvendo empresas e o vínculo societário entre as mesmas. Grilo [Grilo Junior 2010] e Fraga [Fraga 2017] também utilizaram técnicas de mineração de dados para detecção de conluio em licitações. Apesar de identificarem detecção de conluio em licitações, os estudos citados não adentraram no universo de microexpressões, atuando basicamente na busca de padrões frequentes entre licitantes. Balaniuk [Balaniuk et al. 2013] utilizou o classificador Naive Bayes para detectar fraudes em contratos de licitações e apresentou algumas análises baseadas em multicritérios, como a relação entre fornecedores e órgãos. Entretanto, essas análises foram apontadas como features aplicadas ao classificador. Este trabalho pretende avançar na detecção de

fraudes em licitações não só apontando aquelas empresas que fazem parte de um possível conluio mas também identificando e automatizando a detecção das mais variadas formas de camuflar ou dissimular esse comportamento. Essa identificação é realizada através das microexpressões constatadas ao longo do processo.

3. ANOMALIAS EM LICITAÇÕES

Nesta seção descrevemos o problema de detecção de anomalias em licitações, em particular padrões de comportamento anti-competitivos. Um contexto delimita um cenário de auditoria, o qual se caracteriza por um conjunto de licitações determinado pela natureza dos bens ou serviços, pelas entidades participantes, pelo intervalo de tempo, por características das licitações, ou uma combinação dos anteriores.

Uma licitação pode ser vista como uma sequência temporal de eventos realizada por uma ou mais entidades. Tanto a licitação, quanto os eventos, quanto as entidades participantes são caracterizados por atributos que dependem da natureza da licitação. As entidades, no âmbito de compras públicas, são fornecedores e órgãos que adquirem produtos e serviços. Os eventos são as ações de compra, executadas por entidades ou mesmo grupos de entidades.

Licitações anômalas se diferenciam do perfil usual de contratações. Um exemplo de anomalia a prática anti-competitiva, também conhecida por conluio, que se caracteriza por relações pouco usuais entre conjuntos de entidades.

O ponto de partida do nosso problema é um conjunto de entidades E, onde cada entidade e é caracterizada por conjunto de atributos e_p, os quais incluem o conjunto de ações que e pode realizar ou participar. Cada ação realizada por uma ou mais entidades é um evento x, caracterizado pelas entidades participantes x_e, momento no tempo x_t e atributos que caracterizam o evento x_p. Dado um contexto $\mathcal{C}$, o universo de eventos é identificada por $X_{\mathcal{C}}$.

Um padrão é uma expressão lógica de predicados e um escore que quantifica a sua relevância. Cada predicado, por sua vez, é uma expressão relacional contendo instâncias de x_e, x_t ou x_p. O escore pode ser uma combinação de escores dos predicados ou função da expressão como um todo, entre outros.

Dado um conjunto de eventos e o domínio de expressões a ele associado, o problema alvo é determinar os conjuntos de eventos que maximizem o escore dos vários sub-conjuntos possíveis e expressões satisfeitas.

Uma estratégia força-bruta para resolver o problema é enumerar todos os conjuntos de eventos possíveis, assim como o universo de expressões a eles associados, calcular o escore de cada conjunto e expressão, e ordená-los segundo o seu escore, o que é claramente inviável mesmo para cenários com poucos eventos, uma vez que a complexidade da estratégia é $O(2^{|E|} \times X_{\mathcal{C}}^2 \times 2^{|X_c|})$.

Na próxima seção apresentamos uma estratégia gulosa para determinar os padrões desejados e os seus escores.

4. METODOLOGIA

A nossa metodologia se divide em quatro fases que são aplicadas em sequência, descritas a seguir.

4.1 Concorrentes Frequentes

Esta fase determina entidades que frequentemente participam das mesmas licitações. Entidades e podem ser órgãos da Administração Pública Federal que realizam licitações ou fornecedores participantes de uma licitação. As entidades do tipo órgão serão identificadas como e_o e as entidades do tipo fornecedor ou empresas como e_f.

As entidades e_f, durante uma licitação l, irão executar as ações de oferecer lances dentro de uma sequência temporal. Cada pregão será composto de uma série de eventos até que a entidade órgão e_o, representada pelo pregoeiro, encerre o processo no tempo t_l.

A determinação de entidades concorrentes é realizada pelo algoritmo *Apriori* [Agrawal and Srikant 1994]. A entrada do algoritmo é um arquivo contendo as entidades que participaram das licitações, um item de licitação por linha do arquivo. Os conjuntos resultantes são ordenados por valor de *lift* e valores de contratos.

Vale ressaltar que a coocorrência de entidades em um evento, por si só, não poderia ser considerada uma prática anômala, até porque entidades de um mesmo contexto tendem a concorrer em licitações desse contexto. Entretanto, essa participação reiterada e outros atributos decorrentes dos seus eventos podem caracterizar uma anomalia que deve ser avaliada.

4.2 Correlação temporal

A segunda fase é determinar o padrão temporal de ocorrência dos conjuntos de concorrentes frequentes. Desta forma, para cada conjunto, criamos séries ordenadas temporalmente, onde cada valor da série representa o número de itens de licitação na qual o conjunto concorrente participou.

É então calculada a correlação multivariada de cada série e as entidades que apresentam correlação forte ou muito forte[1] são analisadas considerando múltiplos critérios através da instanciação de microexpressões.

4.3 Microexpressões

Definidos os conjuntos de entidades de interesse, com base em co-ocorrências e correlações temporais, instanciamos as microexpressões para todos os conjuntos. Essas microexpressões materializam trilhas de auditoria, que podem ser determinadas com base em experiência pregressa ou legislação.

As microexpressões variam ao longo do tempo entre os diferentes grupos ou de acordo com o contexto abordado entre entidades. Dessa forma, para uma melhor separação entre os diferentes grupos ou entidades, deve-se identificar se o contexto a ser trabalhado se refere a uma obra, a uma prestação de serviço, a compras de quais tipos de objetos ou qualquer outro tipo de divisão de mercado existente.

Quanto maior o número de microexpressões satisfeitas (se binárias) ou a sua magnitude (se contínuas) para os conjuntos de entidades selecionados, maior será a atenção que deve-se dedicar à potencial anomalia encontrada. As microexpressões podem ser associadas a entidades ou licitações e serão descritas a seguir.

4.3.1 *Microexpressões de entidades.*

As microexpressões de entidades estão relacionadas ao conjunto de licitações de uma entidade ao longo do período analisado. São elas:

$\mathcal{E}_{noprim}$: Percentual do valor total dos contratos celebrados nos quais a entidade não foi a primeira colocada.

A licitação consiste de uma sequência de lances submetidos pelas entidades fornecedoras e, findo o prazo da licitação, a entidade que tiver feito o lance de melhor valor é classificada em primeiro lugar. Nesse caso, espera-se que a entidade vencedora celebre o contrato para fornecimento do objeto

[1] correlação forte apresenta coeficiente de pearson entre 0.7 e 0.9. Correlação muito forte apresenta coeficiente de pearson acima de 0.9

da licitação. Caso o contrato não seja celebrado com a vencedora por alguma razão, é convocada a entidade classificada em segundo lugar e assim sucessivamente.

Como objetivo do pregão é comprar pelo menor preço, quando ocorrem casos em que a vencedora, que ofertou melhor preço, não celebra o contrato, a Administração pode acabar contratando por um preço mais elevado, gerando prejuízo ao erário.

Desta forma, para cada uma das entidades fornecedoras analisadas, foi identificada a frequência com que esse comportamento ocorre, ou seja, em todos os contratos que a referida entidade celebrou com a Administração Pública, qual a razão entre o valor total dos contratos em que ela não foi a primeira colocada no certame e o valor total dos contratos celebrados.

$\mathcal{E}_{motdesc}$: Distribuição dos motivos mais frequentes de desclassificação.

Dada a reiteração das desistências ou desclassificações, as atas de pregão foram investigadas com a finalidade de descobrir os principais motivos de desclassificação. Essa verificação é feita de forma manual porque essa informação não consta na base de dados do *SIASG* e sim na ata disponível no site do *Comprasnet*, em formato pdf. Não foram analisadas todas as atas, mas sim uma amostra selecionada a partir das licitações em que as empresas correlacionadas celebraram contrato mas haviam sido classificadas, em geral, depois da décima posição no certame e dos contratos de maior valor. A concentração de motivos de desclassificação foi considerada como indício de anomalia.

$\mathcal{E}_{mercado}$ Percentual de mercado de um fornecedor.

Essa microexpressão analisa o mercado como um todo, dado que o *SIASG* possui licitações dos mais variados produtos e serviços, e essa definição permite reconhecer a real existência de cartéis. Na base de dados não existe um campo que identifique o mercado em questão. Possui apenas o campo de descrição do item licitado ou o campo objeto da licitação. Esses campos são textuais e sem parâmentro obrigatório de preenchimento. Para encontrar o mercado de um contexto, foram selecionadas as entidades que partiparam de qualquer licitação em que pelo menos uma dos fornecedores correlacionados participou ou que tenham em seu objeto as palavras chaves que caracterizam o mercado. Essa análise não exaure o mercado nos contextos de compras de determinados produtos ou prestações de serviços, visto que podem existir licitações em que alguma das empresas correlacionadas não participou e que não tenham as palavras chaves destacadas na descrição do seu item. Entretanto, acreditamos que abarca a maior parte do mercado.

$\mathcal{E}_{acordo}$: Conjuntos de entidades predominantes nas licitações de um órgão.

Uma predominância de fornecedores em determinados órgãos possivelmente significa uma divisão de mercado entre as empresas correlacionadas. Desta maneira, essa microexpressão destaca um presumível acordo entre as empresas para extratificação dos contratos celebrados em órgãos da Administração Pública. As entidades órgãos foram agrupadas por entidades fornecedoras caracterizados pelo valor total e número de contratos por órgão. Os órgãos são então analisados na ordem decrescente da caracterização.

4.3.2 *Microexpressões de licitações.* As microexpressões de licitações são calculadas considerando o conjunto de licitações determinadas pelos conjuntos de entidades concorrentes temporalmente correlacionadas, ou seja, aquelas licitações onde todas as entidades do conjunto participaram de cada licitação do conjunto. São elas:

$\mathcal{L}_{dd}$: Número de entidades desistentes ou desclassificadas.

Uma entidade desclassificada é um fornecedor que não chega a fazer parte da disputa, sendo eliminado na primeira ação, ou seja, no evento em que oferecem o seu primeiro lance, no contexto das entidades analisadas. Geralmente, o próprio edital de licitação aponta o valor estimado e afirma que, serão prontamente desclassificadas as empresas que ofertarem valor maior que aquele valor. Ainda

assim, muitas empresas violam essa regra, apresentam valores acima do estipulado e são prontamente desclassificadas. Entidades desistentes são aquelas que foram vencedoras mas, por algum motivo não celebraram contrato, no contexto das entidades analisadas. Parece plausível, em princípio, que uma empresa seja desclassificada. No entanto, se a frequência com que esse evento acontece e a quantidade de empresas que são desclassificadas ou desistem é significativo, destoa do funcionamento normal de um pregão, representando um possível arranjo ou combinação entre as empresas.

$\mathcal{L}_{perfildd}$: Razão entre os percentuais de desistência ou abstenção de empresas nas licitações selecionadas e no universo de licitações do contexto.

As empresas que mais frequentemente desistem foram ordenadas para examinar como elas se comportam no mercado. Várias delas não possuem contratos homologados quando as empresas analisadas celebram contrato. Ou em alguns casos elas não possuem contrato nenhum celebrado com a administração, participando dos certames sem vencer nenhum deles. Esse fato pode ser indicativo de um possível acordo entre as empresas para disfarçar uma concorrência mas representar, na verdade, um conluio. Essa análise foi realizada verificando-se como essas entidades se comportam em outros certames.

$\mathcal{L}_{sobrepreco}$ Ocorrência e percentual de sobrepreço.

Caso a Administração Pública não celebre o contrato com o primeiro classificado, ela poderá convocar os concorrentes subsequentes, em ordem de classificação, e tentar negociar o preço com os mesmos. Entretanto, essa negociação não é obrigatória. O fornecedor pode manter seu preço original. Esse evento pode culminar na contratação por um preço acima do inicialmente acertado, assinalando um provável sobrepreço. O valor total de sobrepreço foi calculado utilizando o somatório da diferença entre o valor oferecido pela primeira colocada e o valor real de celebração do contrato. Entretanto, caso o contrato seja celebrado com a primera colocada, não será calculado sobrepreço.

4.4 Análise Multicritério

Cada uma das microexpressões indagadas anteriormente parecem apontar indícios de fraude. A combinação de todas elas ou de parte das mesmas gera maior robustez à identificação de conluio, inclusive de um provável cartel. A seleção dessas microexpressões, definindo pesos e características para as mesmas, visa criar um modelo automático de *machine learning* para detecção de fraude.

5. ESTUDO DE CASO

Nesta seção apresentamos um estudo de caso utilizando dados de licitações realizadas por órgãos federais no estado de Minas Gerais, no período de janeiro de 2013 a dezembro de 2018, da base de dados do *SIASG* (banco de dados que contém as informações do *Comprasnet*, [2]). Em particular, consideramos 726057 itens de licitação e 28487 fornecedores.

Para a determinação de concorrentes frequentes, o algoritmo *Apriori* foi executado com os parâmetros finais de 0.001 de suporte e 0.6 de confiança. Os valores de suporte foram baixos dada a pulverização das licitações na base de dados. Os dados extraídos contêm licitações de todos os órgãos do estado de Minas Gerais, nos mais variados contextos. Sendo assim, valores de suporte muito altos poderiam descartar conjuntos relevantes, considerando que normalmente os fornecedores variam de um contexto para outro. Para selecionar os conjuntos de maior potencial de anomalia, foram considerados os valores de *lift* e ordenadas na base de dados de acordo com os maiores valores de contratos. Em seguida, a geração das séries temporais se baseou na data dos itens licitados e da frequência de cada entidade por item nas respectivas datas. Se a entidade concorreu a 10 itens em um mesmo processo de

[2]Comprasnet é um portal de compras com o objetivo de disponibilizar informações sobre licitações e contratações realizadas pelo Governo Federal - https://www.comprasgovernamentais.gov.br/

Table I. Valores e Quantidades Homologados X Homologados em Outras Posições 2008 a 2019 - Brasil

Ent.	$\#l$	$\#i$	$\#i_{prim}$	$\#l_{noprim}$	$\%l_{noprim}$	$Homolog	$Homolog$\#1^o$
A	1071	3219	825	546	66%	R$249.792.951,15	R$179.298.059,90
C	793	4310	420	245	58%	R$235.020.112,88	R$145.002.096,90
L	10015	32135	1548	1244	80%	R$500.960.788,76	R$442.193.399,55
P	746	2214	284	155	55%	R$39.310.272,56	R$26.191.665,02
S	846	2525	143	108	76%	R$89.848.858,41	R$45.137.363,41

licitação, ela terá 10 participações naquela data. Optou-se por utilizar essa forma de contagem para não descartar informações das entidades que concorrem a 1 ou 10 itens em uma mesma licitação. O objetivo da análise de correlação temporal é verificar a evolução dessa relação ao longo dos anos. Nesta etapa foram consideradas as licitações entre janeiro de 2008 e abril de 2019. A correlação temporal foi calculada através do coeficiente de *pearson* multivariado. Algumas entidades que foram indicadas pelo algoritmo *Apriori*, ao serem analisadas através do uso de correlação multivariada, não apontaram resultados muito expressivos ao longo dos anos, mesmo possuindo um valor alto de lift. Outros conjuntos de empresas apresentaram alto índice de correlação multivariada e foram analisados sob a perspectiva de microexpressões. Após a análise das correlações temporais, identificou-se um grupo de 5 entidades que parecem estar altamente correlacionadas e apresentam comportamentos que merecem ser avaliados. Portanto, as microexpressões foram instanciadas considerando-se o contexto delimitado pelas licitações das quais participaram essas 5 entidades. O resultado da microexpressão $\mathcal{E}_{noprim}$ é exibido na Tabela I. Os campos $\#l$ e $\#i$ indicam a quantidade de procedimentos licitatórios que a entidade e participou e a quantidade de itens que e concorreu respectivamente. O campo $\#i_{prim}$ indica a quantidade de itens em que e venceu classificada em primeiro lugar. O campo $\#l_{noprim}$ representa a quantidade de itens em que e celebrou contrato mas não ficou classificada em primeiro lugar. O campo $\%l_{noprim}$ representa, entre as licitações em que a empresa celebrou o contrato, qual o percentual em que ela não ficou em primeiro lugar. Por fim, os campos *$Homolog* e *$Homolog$\#1^o$* significam valor homologado nos contratos em que a empresa celebrou e valor homologado nos contratos em que a empresa celebrou mas não ficou classificada em primeiro lugar, respectivamente. Das licitações em que essas entidades venceram, em média, elas não foram classificadas em primeira colocada em mais de 50%, ou seja, outra empresa ganhou mas foi desclassificada ou desistiu, ocasionando a assinatura do contrato por alguma das empresas do possível cartel. Em alguns casos, esse percentual alcançou 80%.

No que diz respeito às microexpressões $\mathcal{L}_{dd}$ e $\mathcal{L}_{perfildd}$, analisando as empresas desistentes, percebe-se que algumas delas não têm participação em eventos de licitações sem as entidades do cartel estarem participando. Além disso, várias dessas entidades não venceram nenhuma licitação quando as entidades do cartel participaram, mas venceram quando as entidades cartel não estavam presentes ou não foram vencedoras. Em alguns casos, essas empresas chegaram a participar de mais de 100 licitações e não venceram ou celebraram contrato em nenhuma. Com relação à microexpressão $\mathcal{E}_{motdesc}$, os motivos de desclassificação mais frequentes encontrados no contexto analisado, os principais são: entidades que vencem com menor preço mas não apresentam a documentação para celebração do contrato; entidades desclassificadas pelo pregoeiro por apresentar o preço inexequível; e entidades desclassificadas por apresentarem problemas nas planilhas de custos; entidades desclassificadas a pedido delas mesmas. Relativamente a microexpressão $\mathcal{L}_{sobrepreco}$, todas as 5 empresas do cartel analisado apresentaram sobrepreço, sendo que a entidade A apresentou mais de 20 milhões em sobrepreço de acordo com as regras definidas nessa microexpressão.

Considerando a microexpressão $\mathcal{E}_{mercado}$, verificou-se que as entidades A e C possuem os maiores valores de contratos celebrados entre 2008 e 2019, chegando este valor a até 4 vezes os valores de contratos celebrados pelas outras entidades do setor. As entidades L, P e S estão entre as próximas 13 maiores empresas do mercado em questão, considerando também os valores de contratos celebrados. Outra análise de microexpressões que apontou indícios a serem levados em consideração foi a participação das empresas nos órgãos selecionados de acordo com a microexpressão $\mathcal{E}_{acordo}$. As entidades do cartel parecem estar se mantendo nos mesmos órgãos ao longo desses 10 anos, mostrando uma

possível divisão de mercado entre elas.

É interessante notar que a aplicação da metodologia proposta em um caso real permitiu a detecção de um cartel de forma eficiente e assertiva. Maiores detalhes sobre as análises foram omitidos por limitações de espaço ou confidencialidade dos dados.

6. CONCLUSÕES

Este trabalho apresentou uma estratégia de auditoria da participação de empresas em licitações através do uso de mineração de padrões frequentes, correlação multivariada de séries temporais e análise conjugada de multicritérios e detectou não somente possíveis cartéis como microexpressões que explicam a forma de atuação dos mesmos.

Apresentamos também um estudo de caso onde identificamos 5 empresas que, além de apresentarem alta correlação, apresentam comportamento suspeito, pois 50% dos contratos que celebraram não foram as primeiras colocadas. No que se refere a permanência nos órgãos, parecem seguir um padrão, renovando contratos sempre nos mesmos orgãos. A quantidade de empresas desistentes que participam de licitações e não ganham quando as 5 empresas vencem também é alta. Por fim, os motivos de desistências não variam muito e as 5 empresas possuem grandes valores de contratos celebrados e aditivos considerando o mercado analisado em MG.

Em termos de trabalhos futuros, pretende-se automatizar a derivação de microexpressões, assim como o seu processo de análise conjugada. Essas e outras melhorias também serão analisadas no âmbito de outros cenários de compras governamentais.

AGRADECIMENTOS

Os autores agradecem à FAPEMIG, CNPq e CAPES pelo apoio financeiro. Este trabalho também foi parcialmente financiado pelos projetos InWeb, MASWeb, EUBra-BIGSEA, INCT-Cyber, ATMOSPHERE e pela CGU.

REFERENCES

Agrawal, R. and Srikant, R. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*. VLDB '94. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 487–499, 1994.

Balaniuk, R., Bessiere, P., Mazer, E., and Cobbe, P. Corruption risk analysis using semi-supervised naïve bayes classifiers. *International Journal of Reasoning-based Intelligent Systems* vol. 5, pp. 237 – 245, 2013.

Fraga, A. *Detecção de Casos Suspeitos de Fraude em Licitações Realizadas no Município da Paraíba*. Ph.D. thesis, Universidade Federal da Paraíba, Brasil, 2017.

Ghedini Ralha, C. and Sarmento Silva, C. V. A multi-agent data mining system for cartel detection in brazilian government procurement. *Expert Syst. Appl.* 39 (14): 11642–11656, Oct., 2012.

Grilo Junior, T. *Aplicação de Técnicas de Data Mining para Auxiliar o Processo de Fiscalização*. Ph.D. thesis, Universidade Federal da Paraíba, 2010.

Gutflaish, E., Kontorovich, A., Sabato, S., Biller, O., and Sofer, O. Temporal anomaly detection: calibrating the surprise. *CoRR* vol. abs/1705.10085, pp. 1705.10085, 2017.

Hallac, D., Vare, S., Boyd, S. P., and Leskovec, J. Toeplitz inverse covariance-based clustering of multivariate time series data. *CoRR* vol. abs/1706.03161, pp. 1706.03161, 2017.

Siddiqui, M. A., Fern, A., Dietterich, T. G., Wright, R., Theriault, A., and Archer, D. W. Feedback-guided anomaly discovery via online optimization. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. ACM, New York, NY, USA, pp. 2200–2209, 2018.

Song, D., Xia, N., Cheng, W., Chen, H., and Tao, D. Deep r -th root of rank supervised joint binary embedding for multivariate time series retrieval. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. ACM, New York, NY, USA, pp. 2229–2238, 2018.

Tian, K., Zhou, S., Fan, J., and Guan, J. Learning competitive and discriminative reconstructions for anomaly detection. *CoRR* vol. abs/1903.07058, pp. 1903.07058, 2019.

YAGOUBI, D. E., Akbarinia, R., Kolev, B., Levchenko, O., Masseglia, F., Valduriez, P., and Shasha, D. Parcorr: efficient parallel methods to identify similar time series pairs across sliding windows. *Data Mining and Knowledge Discovery* vol. 32, 08, 2018.