

Extração de dados da Web relativos a licitações e contratos públicos para inferência por reconhecimento de padrões estatísticos: estudo de caso.

Cirilo Max Macedo de Moraes e Díbio Leandro Borges

Resumo—Este artigo relaciona técnicas para extração de dados da Web relativos a licitações e contratos públicos para inferência por reconhecimento de padrões estatísticos, utilizando a linguagem R, e apresenta um estudo de caso para identificação da correlação entre o maior número de contratos realizados por uma unidade administrativa com uma mesma empresa e o número total de contratos celebrados por esta unidade administrativa. As informações foram extraídas do sítio da internet www.dados.gov.br.

Palavras-Chave— extração de dados da Web, contratos públicos, reconhecimento de padrões, dados abertos, dados governo.

Abstract—This article lists techniques for extracting Web data relating to tenders and public contracts for statistical inference for pattern recognition, using the R language, and presents a case study to identify the correlation between the maximum number of public contracts from an administration unity with the same company and the total number of contracts from this administration unity. The information was extracted from the Web site www.dados.gov.br.

Keywords—Web data extraction, public contracts, pattern recognition, open data, government data.

I. INTRODUÇÃO

Identificar padrões em licitações e contratos públicos fraudulentos é de fundamental importância para a melhor utilização dos recursos humanos dos órgãos de controle governamentais. A enorme quantidade de licitações públicas, juntamente com a escassez de pessoal nos órgãos de controle, faz com que as verificações dos contratos sejam realizadas em um pequeno percentual. O critério de escolha dos contratos a serem auditados é realizado por denúncias, amostragem aleatória, montante dos recursos envolvidos, ausência de prestação de contas ou outros. Tais critérios, por sua vez, implicam na não identificação de irregularidades em muitos contratos analisados e permitem que o restante dos contratos não sofra nenhum tipo de auditoria externa.

A fim de criar um critério alternativo, uma análise automática em todos os contratos realizados na administração pública poderia identificar padrões por tipos de objeto e apontar aqueles com maior possibilidade de ocorrência de irregularidades.

Além disso, o surgimento de nova modalidade de licitação como o Regime Diferenciado de Contratação – RDC, que flexibiliza a obrigatoriedade de existência prévia de projeto executivo e orçamento detalhado para contratação de obras e serviços de engenharia, torna ainda mais necessário o

conhecimento prévio de padrões estatísticos de contratos similares.

Como fonte de dados para o reconhecimento de padrões, seriam utilizados os dados públicos disponíveis na Internet, disponibilizados pelo governo federal, que elaborou uma ferramenta para isto. Esta ferramenta, conhecida como Portal Brasileiro de Dados Abertos, apresenta no sítio www.dados.gov.br informações sobre convênios, compras governamentais e outros. Os dados são disponibilizados através de APIs (*Application Programming Interfaces*).

Para exemplificar a possibilidade de inferência estatística nestes dados públicos, o artigo apresenta um estudo de caso mostrando os resultados obtidos mediante regressão linear e sua correspondência à hipótese formulada.

No estudo de caso, pretende-se verificar se existe uma correlação entre o número de contratos celebrados por uma mesma unidade administrativa com uma mesma empresa e o número total de contratos desta unidade administrativa, e se esta correlação varia de acordo com a modalidade de licitação.

A hipótese é que como as licitações públicas buscam celebrar o contrato mais vantajoso para a União e, para tal, devem respeitar, dentre outros, os princípios da moralidade e isonomia, então o número total de contratos celebrados entre uma unidade administrativa e uma mesma empresa deveria supostamente ser proporcional ao número total de contratos celebrados por uma unidade administrativa, preservando uma mesma correlação, independentemente da modalidade de licitação escolhida.

A apresentação do artigo é realizada, inicialmente, mostrando a fonte de dados disponibilizada pelo governo federal e utilizada para o estudo de caso. Em seguida são mostradas algumas técnicas, na linguagem R, utilizadas para a obtenção de dados na internet. Posteriormente é apresentada uma descrição resumida das principais modalidades de licitação. Após a abordagem dos principais assuntos envolvidos no problema analisado, inicia-se a explicação dos procedimentos de extração, análise dos dados e resultados obtidos no estudo de caso. Por fim são apresentadas as conclusões e possibilidades de futuras pesquisas.

II. ESTUDO DE CASO

A. Fonte de Dados

Para a realização do estudo de caso, foram utilizados dados extraídos do portal de dados abertos brasileiro, cujas características são descritas em seguida.

O Brasil tornou-se membro do *Open Government Partnership*, uma iniciativa multinacional para adoção mundial do *Open Government Data* (OGD) em setembro de 2011 [2]. O portal de dados abertos brasileiro foi lançado em 2012 e

Cirilo Max Macedo de Moraes, Mestrado Profissional em Engenharia Elétrica com ênfase em Informática Forense e Segurança da Informação - Departamento de Engenharia Elétrica, Universidade de Brasília, Campus Universitário Darcy Ribeiro, Prédio CIC/EST, Asa Norte, 70910-900 - Brasília, DF - Brasil, e-mail: cirilomax@gmail.com. e Díbio Leandro Borges, Departamento de Ciência da Computação, Universidade de Brasília, Campus Universitário Darcy Ribeiro, Prédio CIC/EST, Asa Norte, 70910-900 - Brasília, DF - Brasil, e-mail: dibio@unb.br.

disponibilizado no sítio da internet www.dados.gov.br [3]. O desenvolvimento do projeto é parte da Infraestrutura Nacional de Dados Abertos (INDA) e tem por objetivo a centralização de acesso ao maior conjunto possíveis de dados públicos governamentais. Em abril de 2015, estavam disponíveis para a consulta 859 conjuntos de dados. Nestes conjuntos, estão disponíveis vários tipos de dados, como da saúde suplementar, do sistema de transporte, de segurança pública, indicadores de educação, gastos governamentais, processo eleitoral e outros.

Como fonte para este estudo de caso, foi utilizado o conjunto de dados de compras públicas do governo federal. Tais dados abertos são do Sistema Integrado de Administração e Serviços Gerais – SIASG, onde se operacionalizam as compras do Governo Federal. Os dados estão disponíveis através de uma API de Dados Abertos em versão beta. Por meio desta API é possível acessar dados dos fornecedores, do catálogo de materiais, do catálogo de serviços, das licitações e dos contratos.

O acesso aos dados é feito através de URLs com recursos Web nos formatos XML, JSON, CSV e HTML. É possível realizar em cada consulta uma série de parâmetros de filtro que devem compor a URL. Para acessar os dados da API é necessário informar o formato de resposta (HTML, XML, JSON ou CSV) e qual informação é desejada[8]. No processo de extração dos dados escolheu-se o formato JSON como formato de resposta.

Para o objetivo do estudo de caso, as informações necessárias seriam os contratos celebrados em cada unidade administrativa (UASG) e as respectivas empresas contratadas. Tais informações podem ser obtidas utilizando-se o método de consulta Contratos disponível no módulo Contratos da API.

A API possibilita a utilização de vários parâmetros na consulta. Como cada consulta fornece informações de apenas 500 contratos, faz-se necessária a utilização do parâmetro *offset*. O parâmetro *offset* corresponde à quantidade de registros ignorados a partir do início da lista de resultados ordenados pelo ID.

Suspeitando-se que a modalidade da licitação pudesse influenciar no número de contratos realizados com uma mesma empresa, procurou-se filtrar os dados por modalidade de licitação. Para tal faz-se necessária a utilização do parâmetro modalidade. O parâmetro modalidade corresponde ao código da modalidade de licitação.

Desta forma, a API filtra os contratos por modalidade de licitação mediante a utilização das seguinte URLs:

<http://compras.dados.gov.br/contratos/v1/contratos.json?modalidade=01>, para contratos na modalidade Convite.

<http://compras.dados.gov.br/contratos/v1/contratos.json?modalidade=02>, para contratos na modalidade Tomada de Preços.

<http://compras.dados.gov.br/contratos/v1/contratos.json?modalidade=03>, para contratos na modalidade Concorrência.

<http://compras.dados.gov.br/contratos/v1/contratos.json?modalidade=05>, para contratos na modalidade Pregão.

<http://compras.dados.gov.br/contratos/v1/contratos.json?modalidade=06>, para contratos na modalidade Dispensa de Licitação.

<http://compras.dados.gov.br/contratos/v1/contratos.json?modalidade=07>, para contratos na modalidade Inexigibilidade de Licitação.

B. R-Project

Antes da descrição dos procedimentos para obtenção dos dados utilizados no estudo de caso, nesta seção são apresentadas algumas maneiras de extração de dados da Web.

A extração de dados da Web pode ser realizada em vários níveis, desde a análise de baixo nível dos formatos HTML / XML / JSON até a chamada final de funções simples implementadas por vários pacotes de conveniência, que podem por exemplo fazer uso internamente de especificações de API [5].

A linguagem R fornece uma gama de funcionalidades e recursos que possibilitam a extração e análise de dados da Web de uma maneira rápida e simples. Sua utilização cresceu bastante no mundo nos últimos anos. Por este motivo foi escolhida para a realização do estudo de caso.

R é uma linguagem de programação livre e um ambiente de *software* que fornece uma grande variedade de técnicas estatísticas e gráficas.

A CRAN (*Comprehensive R Archive Network*) é uma rede de FTP e servidores Web ao redor do mundo que armazena versões do código e documentação do R idênticos e atualizados.

Dentro da comunidade R muitos projetos contribuíram para fornecer infraestrutura que permitissem a interação do R com a Web. Recentemente, uma grande quantidade de pacotes de correspondência com tecnologias Web tem sido desenvolvida e é agora coletada, organizada e estruturada em uma vista de tarefas de tecnologias Web CRAN disponível no <http://cran.r-project.org/web/views/WebTechnologies.html> [5].

Os pacotes do R que permitem a extração de dados da Web em baixo nível são XML, RCurl e rjson/RJSONIO/jsonlite. O pacote XML contém funções para análise XML e HTML e suporta XPath para pesquisa em XML. Uma importante função em XML para extrair dados de uma ou mais tabelas HTML é `readHTMLTable()`. O pacote RCurl faz uso da biblioteca libcurl para transferência de dados usando vários protocolos, permite compor solicitações HTTP e fornece convenientes funções para busca de URLs, métodos get/post e outros, além de processar os resultados retornados pelo servidor Web. JSON (JavaScript Object Notation) é um formato de intercâmbio de dados que se tornou o mais comum da Web. Os pacotes rjson, RJSONIO, e jsonlite convertem objetos R em objetos JSON e vice-versa[5].

C. Modalidades de Licitações

No estudo de caso realizado, procurou-se distinguir os contratos celebrados por diferentes modalidades de licitação. De uma maneira sucinta são relacionadas as principais modalidades nesta seção.

A lei nº 8.666, de 21 de junho de 1993, instituiu as normas para licitação e contratos da administração pública. De acordo com a referida lei, existem várias modalidades de licitação, dentre as quais, carta-convite, tomada de preços e concorrência. A lei nº 10.520 de 17 de julho de 2002 instituiu a modalidade de licitação denominada pregão para a aquisição de bens e serviços comuns, sendo posteriormente regulamentada na forma eletrônica através do Decreto nº 450, de 31 de maio de 2005.

A dispensa de licitação foi autorizada em casos especiais de compra, sem desrespeitar os princípios da moralidade e isonomia, devendo-se limitar à aquisição de bens e serviços indispensáveis ao atendimento da situação de emergência ou dos casos descritos no artigo 24 da lei 8.666/93, dentre eles contratação de pequeno valor.

A inexigibilidade de licitação ocorre devido à inviabilidade de competição em situações descritas no artigo 25 da lei 8.666/93, como existência de fornecedor exclusivo, de empresa ou pessoa física com notória especialização, profissional artista, dentre outros.

Antes da abordagem dos dados obtidos é necessário apresentar algumas técnicas para extração de dados da Web.

Como exemplos de diferenças entre as citadas modalidades, pode-se citar o prazo entre o aviso de publicação do edital e o recebimento das propostas e os valores máximos do contrato permitido para cada modalidade.

Quanto à antecedência de publicação do edital para licitações que não envolvam técnica ou empreitada integral, os prazos mínimos são:

- a) convite – 5 dias úteis;
- b) tomada de preços – 15 dias;
- c) concorrência – 30 dias;
- d) pregão – 8 dias úteis;

Quanto ao valor do contrato, há distinção entre os objetos. Para obras e serviços de engenharia, o valor máximo de cada modalidade é maior que para compras e serviços. Os valores máximos para obras e serviços de engenharia e para compras e serviços, respectivamente, são mostrados a seguir:

a) convite - até R\$ 150.000,00 (cento e cinquenta mil reais) e R\$ 80.000,00 (oitenta mil reais) ;

b) tomada de preços - até R\$ 1.500.000,00 (um milhão e quinhentos mil reais ou R\$ 650.000,00 (seiscentos e cinquenta mil reais);

c) concorrência - acima de R\$ 1.500.000,00 (um milhão e quinhentos mil reais) ou de R\$ 650.000,00 (seiscentos e cinquenta mil reais);

d) pregão – sem limite de valor.

Deve-se ressaltar que a lei 12.462, de 5 de agosto de 2011, instituiu o Regime Diferenciado de Contratações Públicas – RDC. Este regime foi inicialmente previsto para aplicação nas licitações e contratos necessários para os grandes eventos como Copa do Mundo, Copa das Confederações e Olimpíadas. Atualmente, muitas outras leis ampliaram o escopo do objeto do RDC. Neste modelo, a seleção do fornecedor se processa pela escolha da menor proposta realizada por meio de lances públicos e diferencia-se das modalidades de licitação, citadas anteriormente, dentre outros fatores, pelo fato de que os concorrentes não conhecem o orçamento, nem possuem o projeto executivo da obra, mas sim, no caso da contratação integrada, um projeto básico que caracteriza a obra com base nas indicações dos estudos técnicos preliminares.

No estudo de caso realizado foram utilizados somente os contratos relacionados às modalidades de licitação: carta-convite, tomada de preços, concorrência, pregão, dispensa de licitação e inexigibilidade de licitação.

D. Extração de Dados

Nesta seção é apresentada a metodologia utilizada para a extração de dados do estudo de caso.

Para a extração de dados foram utilizados os pacotes RCurl e RJSONIO.

Utilizou-se para consulta a seguinte URL descrita anteriormente:

"http://compras.dados.gov.br/contratos/v1/contratos.json"

Para fornecimento dos parâmetro de consulta e obtenção do objeto JSON, utilizou-se a seguinte função:

getForm(url, .params = list (modalidade = mod, offset = as.integer(off))), onde mod e off representam os valores dos parâmetros de consulta.

Para transformação do objeto JSON em objeto R, utilizou-se a função fromJSON().

Para a obtenção dos dados em objeto R, foi necessária também a implementação de procedimentos para padronizar suas estruturas de dados e automatizar a consulta em todas as páginas correspondentes à modalidade de licitação desejada. Tais procedimentos não estão descritos neste artigo.

Ao final do processo de extração dos dados, foram obtidos 6 *dataframes*, correspondentes aos contratos constantes do portal de dados abertos brasileiro, divididos por modalidade de licitação (Tabela I).

TABELA I- DADOS EXTRAÍDOS

Dataframe	Modalidade de Licitação	Número Contratos
1	carta-convite	34.009
2	tomada de preços	39.394
3	concorrência	24.035
4	pregão	219.647
5	dispensa de licitação	89.478
6	inexigibilidade	45.378
TOTAL		451.941

Cada Dataframe era constituído por 17 campos contendo as seguintes informações (Tabela II):

TABELA II – CAMPOS DOS DATAFRAMES

Nº	Campo	Descrição
1	identificador	Identificador do Contrato
2	uasg	Campo de seis dígitos que indica o código da UASG contratante.
3	modalidade_licitacao	Número e o ano da licitação que originou a contratação.
4	numero_licitacao	Número do aviso da licitação que originou a contratação.
5	tipo_contrato	Tipo de Contrato.
6	licitacao_associada	Referência à licitação que originou a contratação.
7	origem_licitacao	Origem da licitação que gerou o contrato: Preço praticado(SISPP) ou Registro de preço(SISRP).
8	numero	Campo seguido pelo número do contrato, seguido do respectivo ano.
9	objeto	Descrição do objeto, a partir de uma descrição de item/serviço informada.
10	numero_aditivo	Quantidade de termos aditivos de um contrato.
11	numero_processo	Número do processo de contratação.
12	cnpi_contratada ou cpf_contratada	CNPJ da empresa contratada ou CPF da contratada, varia de acordo com a modalidade de licitação
13	dataassinatura	Data de assinatura do contrato.
14	fundamento_legal	Fundamento legal do processo de contratação.
15	data_inicio_vigencia	Data de início de vigência dos contratos.
16	data_termino_vigencia	Data de término de vigência dos contratos.
17	valor_inicial	Valor inicial do contrato.

E. Análise dos Dados

Após a extração dos dados passou-se ao tratamento e análise dos mesmos.

Inicialmente identificaram-se os campos necessários para a busca dos padrões de acordo com a descrição da hipótese, ou seja, o CNPJ ou CPF do fornecedor contratado e a unidade administrativa contratante (UASG).

Em seguida, para cada *dataframe*, elaborou-se um novo *dataframe* compreendendo apenas os campos UASG, CNPJ ou CPF contratado (conforme modalidade de licitação e o número de contratos celebrados entre os dois. Para tal utilizou-se a seguinte função R:

```
ddply(dataframe, c("uasg", "cnpj_contratada"), summarise,
      ContrPorEmpresa=length(modalidade_licitacao)
```

),
onde *ContrPorEmpresa* passou a ser o campo do novo *dataframe* que informa a quantidade de contratos celebrados entre uma unidade administrativa e uma mesma empresa.

Finalmente elaborou-se um último *dataframe*, a partir do anterior, compreendendo o campo UASG e criando os campos *NumEmpresas* (número de empresas contratadas por UASG), *NumContratos* (número total de contratos celebrados por cada UASG) e *MaiorNumContMesmaEmpresa* (maior número de contratos que cada UASG celebrou com uma mesma empresa). Para tal foi utilizada a seguinte função R:

```
ddply(dataframe1, c("uasg"), summarise,
      NumEmpresas = length(ContrPorEmpresa),
      NumContratos=sum(ContrPorEmpresa),
      MaiorNumContMesmaEmpresa=max(ContrPorEmpresa)
)
```

Cada *dataframe* final, relativo a cada modalidade de execução, obteve os seguintes dados (Tabela III):

TABELA III – DADOS FINAIS PARA ANÁLISE

Dataframe	Modalidade de Licitação	Número Uasgs	Número Contratos	Número Empresas	M. Num. Cont. Mesma Empr.
1	carta-convite	1.773	34.009	24.055	43
2	tomada de preços	1.781	39.394	25.343	50
3	concorrência	1.247	24.035	13.217	43
4	pregão	2.429	219.647	131.301	61
5	dispensa de licitação	2.590	89.478	49.729	599
6	inexigibilidade	2.160	45.378	24.626	207

Para a verificação da correlação entre o número de contratos de uma unidade administrativa (uasg) e o máximo de contratos celebrados com uma mesma empresa por essa unidade administrativa, foram estudadas as formas de regressão linear e não-linear.

Para a regressão linear, buscou-se encontrar os coeficientes que produzissem a melhor aproximação, utilizando-se o critério dos mínimos quadrados, conforme a seguinte função de regressão:

$$Y \approx \beta_0 + \beta_1 X$$

$$\text{MaiorNumContMesmaEmpresa} \approx \beta_0 + \beta_1 \text{NumContratos}$$

Verificando-se a regressão linear para os contratos da modalidade de licitação carta-convite, obteve-se os seguintes coeficientes (Fig.1):

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.496936   0.052235  28.66  <2e-16 ***
NumContratos 0.054072   0.001238  43.69  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.959 on 1771 degrees of freedom
Multiple R-squared:  0.5187, Adjusted R-squared:  0.5184
F-statistic: 1909 on 1 and 1771 DF, p-value: < 2.2e-16
```

Fig.1 – Coeficientes da regressão linear na Modalidade Carta-convite

Percebe-se que o valor p é praticamente nulo. Além disso os coeficientes são bem maiores que o desvio padrão, levando à obtenção de valores t grandes. Tais condições afastam a hipótese dos coeficientes β_0 e β_1 serem nulos.

Além disso, obtém-se um coeficiente de determinação $R^2 = 0,51$. Tal coeficiente indica a proporção de variabilidade presente nas variações da variável resposta Y que é explicada pela variável independente X no modelo de regressão.

Assim os coeficientes usados no modelo de regressão linear, para os contratos da modalidade carta-convite, foram:

$$\beta_0 = 1.497$$

$$\beta_1 = 0.054$$

Procurando-se agora verificar se um modelo não linear se ajustaria melhor, realizou-se a comparação da regressão linear com funções polinomiais até o grau 5,:

$$Y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5$$

```
> anova(lmMnunemp.1, lmMnunemp.2, lmMnunemp.3, lmMnunemp.4, lmMnunemp.5)
Analysis of Variance Table

Model 1: MaiorNumContMesmaEmpresa ~ NumContratos
Model 2: MaiorNumContMesmaEmpresa ~ poly(NumContratos, 2)
Model 3: MaiorNumContMesmaEmpresa ~ poly(NumContratos, 3)
Model 4: MaiorNumContMesmaEmpresa ~ poly(NumContratos, 4)
Model 5: MaiorNumContMesmaEmpresa ~ poly(NumContratos, 5)

  Res.Df    RSS Df Sum of Sq    F      Pr(>F)
1     1771 6797.5
2     1770 5860.8   1    936.68 294.957 < 2.2e-16 ***
3     1769 5763.0   1     97.84  30.810 3.277e-08 ***
4     1768 5693.2   1     69.77  21.969 2.983e-06 ***
5     1767 5611.4   1     81.84  25.770 4.251e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig.2 – Comparação entre polinômios de diferentes graus.

Pelos valores p e F, percebe-se que os modelos de 3º, 4º e 5º graus apresentam um ajuste melhor que a regressão linear simples. Para o modelo com polinômio de 5º grau, o coeficiente de determinação, R^2 , é igual a 0,60, ou seja, superior ao da regressão linear simples (Fig.2). A Figura 3 a seguir apresenta os diferentes modelos no gráfico:

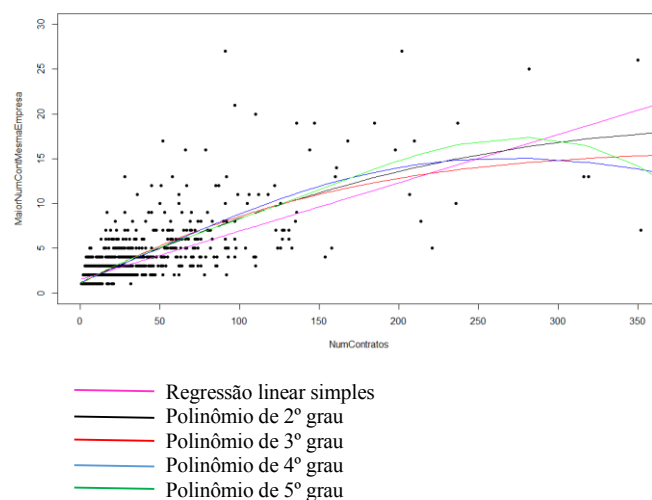


Fig.3 – Modelos de ajuste.

No entanto, devido à proximidade com a regressão linear simples no intervalo de números de contrato contido entre 1 e 150 contratos (maior concentração de unidades administrativas), optou-se pela realização da comparação entre as diversas modalidades de licitação mediante a utilização da regressão linear simples.

A partir dos dados obtidos, foram elaborados os seguintes gráficos para cada modalidade de licitação (Figuras 4 a 9), onde a linha central representa a função de regressão e as linhas verdes as margens superior e inferior de acordo com o nível de tolerância.

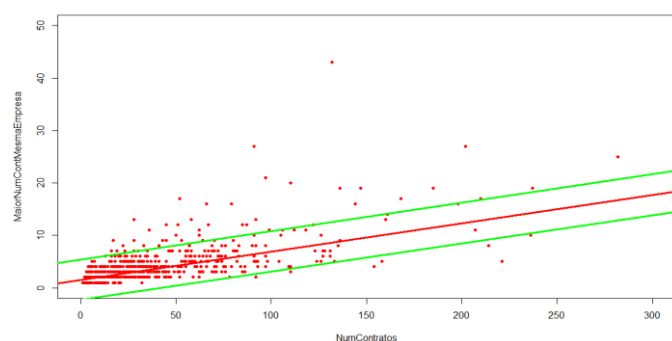


Fig.4 – Modalidade Carta-convite

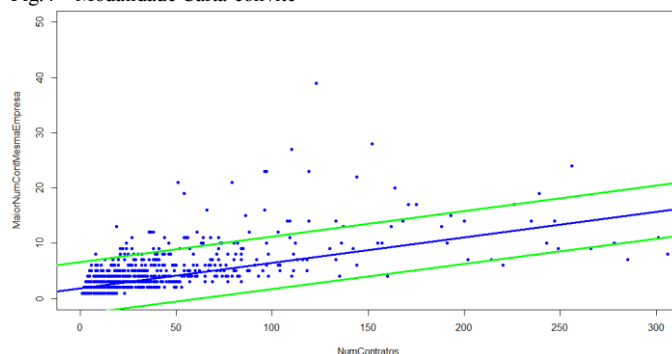


Fig.5 – Modalidade tomada de preços

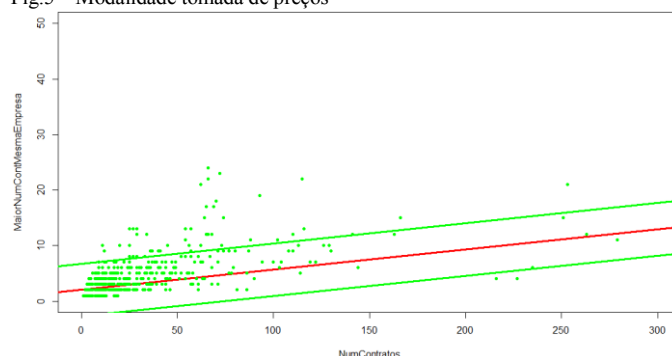


Fig.6 – Modalidade concorrência

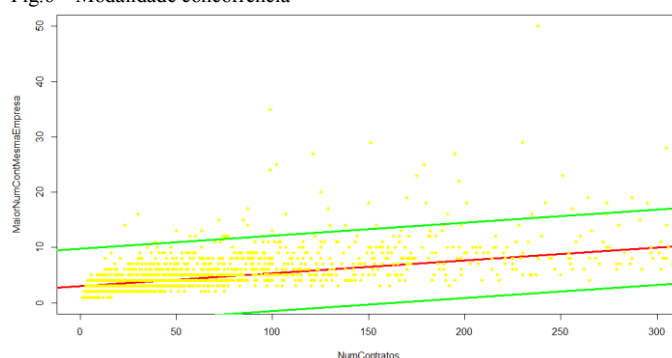


Fig.7 – Modalidade pregão

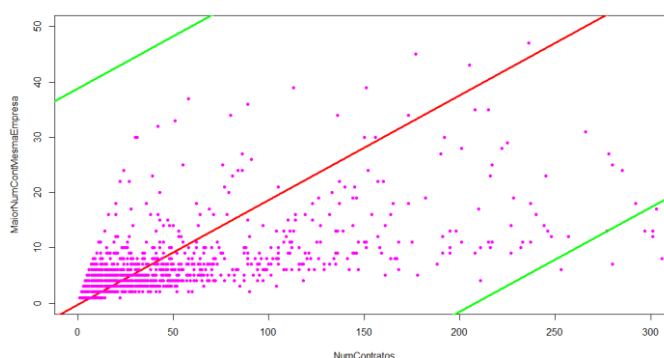


Fig.8 – Modalidade dispensa de licitação

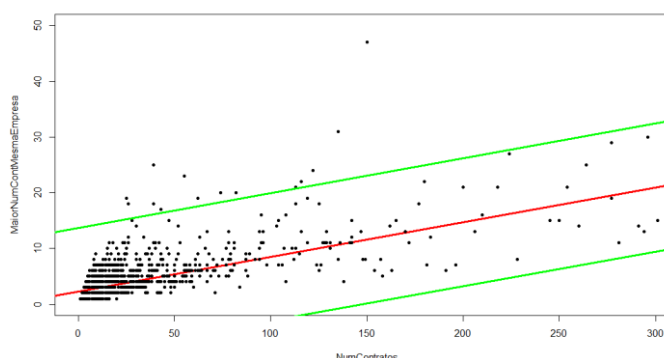


Fig.9 – Modalidade inexigibilidade de licitação

Pelos gráficos anteriores pode-se perceber, para todas as modalidades de licitação, a existência de uma correlação entre o maior número de contratos celebrados com uma mesma empresa e o número total de contratos da unidade administrativa. Ressalta-se que na modalidade dispensa de licitação existe uma dispersão maior.

Comparando-se agora todas as funções de regressão de acordo com as modalidades de licitação obtêm-se o seguinte gráfico (Figura 10):

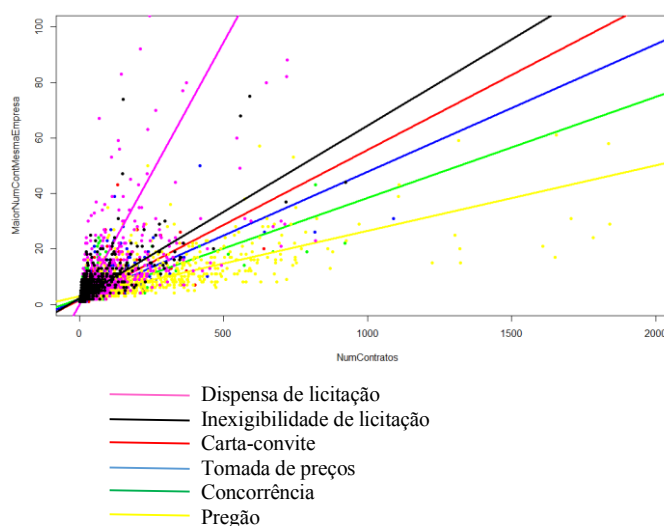


Fig.10 – Comparação entre as modalidades de licitação

F. Resultados Obtidos

A partir da análise da correlação entre os contratos celebrados com uma mesma empresa e o número de contratos celebrados por uma unidade administrativa foi possível estabelecer uma correlação entre as mesmas.

Verificou-se que o número de contratos celebrados com uma mesma empresa possui uma correlação diferente com número

total de contratos da unidade administrativa, variando de acordo com a modalidade de licitação escolhida.

Tal variação pode ser explicada pelas diferentes maneiras de seleção da empresa a ser contratada. Na dispensa de licitação e inexigibilidade de licitação não existe certame, sendo a empresa selecionada diretamente pela unidade administrativa e, portanto, mais fácil de haver uma possível preferência. A modalidade de licitação carta-convite prevê a disputa entre três empresas escolhidas livremente pela administração, o que também diminui a possibilidade de direcionamento, mas ainda possibilita uma certa preferência. Nas demais modalidades de licitação, a exigência de divulgação do edital aumenta, sendo, na modalidade pregão, a disputa realizada através da internet, o que dificulta o direcionamento pela unidade administrativa.

Por outro lado, as modalidades de licitação são realizadas de acordo com o valor, sendo, de uma maneira geral, a dispensa de licitação realizada para os contratos de menor valor e a concorrência para os contratos de maior valor. Outra explicação para os diferentes modelos de regressão poderia ser o aumento do interesse das empresas em licitações de maior valor e consequentemente maior distribuição dos contratos.

Observa-se, assim, que quanto maior a exigência de divulgação do edital de licitação e quanto maior o valor do contrato (modalidade de licitação) menor é a correlação entre o número máximo de contratos celebrados com uma mesma empresa e o número total de contratos celebrado com uma mesma unidade administrativa.

O estudo encontra-se em caráter preliminar e não foram exauridas todas as formas de ajustes possíveis para caracterização do melhor modelo. Da mesma forma, muitas outras informações dos contratos poderiam ser utilizadas para formação deste modelo, como valor do contrato, data de assinatura e outros. No entanto pode-se destacar que é possível inferir um resultado a partir de reconhecimento de padrões estatísticos.

Tal inferência poderia contribuir para a identificação de possíveis irregularidades na contratação, pois a existência de valores muito superiores aos padrões obtidos poderia levantar a suspeita de ocorrência de favorecimento ou direcionamento de licitação. Um tratamento mais refinado nos dados obtidos poderiam, assim, apontar para os órgãos de controle quais unidades administrativas e empresas mereceriam uma análise pormenorizada dos contratos, por exemplo.

III. CONCLUSÕES

A partir da extração de dados da Web relativos a licitações e contratos públicos é possível obter inferências por reconhecimento de padrões estatísticos.

Como importante fonte de dados relacionados a contratos e licitações para obtenção de padrões estatísticos, pode-se utilizar o portal de dados abertos brasileiro, www.dados.gov.br. Além de licitações e contratos, tal portal fornece uma grande quantidade de informações relacionadas a vários tipos de dados como da saúde suplementar, do sistema de transporte, de segurança pública, indicadores de educação, gastos governamentais, processo eleitoral e outros.

Em um estudo de caso preliminar para obter inferências por reconhecimento de padrões estatísticos entre o número de contratos celebrados por uma unidade administrativa com uma mesma empresa e o número total de contratos celebrados por esta unidade administrativa, foi possível estabelecer uma

correlação entre os mesmos. Com a elaboração de um modelo de regressão linear, o estudo de caso explicitado neste artigo, utilizando a linguagem R para extração de dados relacionados a licitações e contratos do portal de dados abertos brasileiros, comprovou que quanto maior a exigência de divulgação de dados do edital e quanto maior o valor do contrato, menor é a correlação entre o número de contratos celebrados entre uma unidade administrativa e uma mesma empresa e o número total de contratos desta unidade.

O estudo de caso abordou apenas três variáveis de dezessete variáveis extraídas dos contratos. Variáveis como valor do contrato, data de assinatura, número de aditivos e outros não fizeram parte do modelo apresentado.

Muitas outras informações, correlações e padrões entre licitações e contratos, não abordados neste artigo, podem ser objeto de estudos.

A partir do conhecimento destes padrões estatísticos, a observância de grandes desvios poderia apontar aos órgãos de controle a suspeita de possíveis irregularidades.

Acrescenta-se que o surgimento de nova modalidade de licitação, como o Regime Diferenciado de Contratação – RDC, que flexibiliza a obrigatoriedade de existência prévia de projeto executivo e orçamento detalhado para contratação de obras e serviços de engenharia, torna ainda mais necessário o conhecimento prévio de padrões estatísticos de contratos similares.

AGRADECIMENTOS

Aos professores do mestrado profissional em engenharia elétrica da UNB pelas orientações, à família e aos colegas pelo constante incentivo e motivação.

REFERÊNCIAS

- [1] MUNZERT, Simon et al. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons, 2014.
- [2] HASTIE, Trevor et al. An Introduction to Statistical Learning with Applications in R. 2013.
- [3] DOS SANTOS BRITO, Kellyton et al. Brazilian government open data: implementation, challenges, and potential opportunities. In: Proceedings of the 15th Annual International Conference on Digital Government Research. ACM, 2014. p. 11-16.
- [4] SEERS, Blake M.; SHEARS, Nick T. New Zealand's Climate Data in R—An Introduction to cliffo. 2015.
- [5] MAIR, Patrick; CHAMBERLAIN, Scott. Web Technologies Task View. A peer-reviewed, open-access publication of the R Foundation for Statistical Computing, p. 178, 2014.
- [6] VARELA, Sara et al. rAvis: an R-package for downloading information stored in Proyecto AVIS, a citizen science bird project. PloS one, v. 9, n. 3, p. e91650, 2014.
- [7] COSTA, Jefferson de J. et al. Uma análise da qualidade dos dados relativos aos boletins de ocorrências das rodovias federais brasileiras para o processo de Mineração de Dados, 2013.
- [8] ———, “Dados Abertos Governamentais” (2012). [Online]. Disponível em: <http://www.governoeletronico.gov.br/acoes-e-projetos/Dados-Abertos>. Acessado em: 27 abr. 2015.
- [9] Karin Breitman, Percy Salas, Marco Antonio Casanova, Daniel Saraiva, Vinicius Gama, Jose Viterbo, Regis Pires Magalhães, Ednylton Franzosi, Miriam Chaves, "Open Government Data in Brazil", IEEE Intelligent Systems, vol.27, no. 3, pp. 45-49, May-June 2012, doi:10.1109/MIS.2012.25.
- [10] Abu-Mostafa, Y; Magdon-Ismael, M. & Lin, H. Learning from Data, AMLBook, EUA, 2012.