Eixo temático:

Metodología en ciencia política

TOGARY:

Um *software* para coleta automatizada de dados sobre corrupção no Brasil[Ω*]

Lucas Silva (UFPE)
lucas.eosilva@ufpe.br

Dalson Figueiredo Filho (UFPE)
dalson.figueiredofo@ufpe.br

Enivaldo Rocha (UFPE)
enivaldocrocha@gmail.com

[Ω] Trabalho apresentado no VIII Congresso Latinoamericano de Ciência Política organizado pela Associação Latinoamericana de Ciência Política (ALACIP), realizado na Pontifícia Universidade Católica do Peru, Lima, entre 22 a 24 de julho de 2015. Agradecemos à Universidade Federal de Pernambuco (UFPE) pelo apoio institucional e ao *Berkeley Initiative for Transparency in the Social Sciences* (BITTS) <http://www.bitss.org>. Esse trabalho conta ainda com o apoio logístico do Grupo de Métodos de Pesquisa em Ciência Política: <dgp.cnpq.br/dgp/espelhogrupo/2341746722315920>. A versão em inglês desse *paper* conta com a participação de Nicole Janz (Universidade de Cambridge). Os materiais para replicação estão disponíveis em: <http://dx.doi.org/10.7910/DVN/27787>. A última atualização do TOGARY foi realizada em 14 de março de 2015. Mais informações sobre como o *software* funciona podem ser encontradas em: <https://www.youtube.com/watch?v=oyNhJ80Mut8>. Eventuais erros são monopólio dos autores.

RESUMO

TOGARY é uma ferramenta computacional para coleta automatizada de informações e disponibiliza o maior banco de dados de corrupção observada no mundo. O programa extrai as informações das sentenças judiciais de improbidade administrativa no *site* do Conselho Nacional de Justiça (CNJ) e exporta os dados em formato de planilha de trabalho. Nossa motivação para criar o *software* é a limitada transparência institucional do CNJ uma vez que os dados originais apenas estão disponíveis para consulta caso-a-caso, o que inviabiliza a coleta manual. A atualização mais recente do TOGARY tem informações para mais de 15.000 casos de corrupção e outros delitos julgados nas esferas estadual, federal, eleitoral e superior no Brasil entre 1992 e 2014. Utilizamos análise descritiva e espacial para ilustrar os dados extraídos pelo TOGARY. Com este trabalho, esperamos difundir a utilização de procedimentos automatizados de coleta de dados em pesquisas empíricas em Ciência Política. Além disso, esperamos expandir a utilização da ciência para combater a corrupção no Brasil.

Palavras-chave: coleta automatizada de dados; transparência institucional; corrupção.

ABSTRACT

TOGARY is a tool for automated data collection and provides the largest open dataset on observed corruption worldwide. The program extracts information on administrative improbity judicial cases from the Brazilian National Council of Justice (CNJ) website and exports it as spreadsheet format. Our motivation to create this software is the lack of institutional disclosure from CNJ since original data is only available for case-by-case inquiry which makes manual collection infeasible. The latest TOGARY update has information on more than 15.000 corruption cases and other crimes judged by Brazilian subnational, national, electoral and superior courts between 1992 and 2014. We use descriptive statistics and spatial analysis to examine data collected by TOGARY. With this paper, we hope to diffuse the application of automated data procedures in Political Science research. Additionally, we hope to advance the use of science to fight corruption in Brazil.

Keywords: automated data collection; institutional transparency; corruption.

INTRODUÇÃO

A quantidade de informações disponíveis na internet oferece uma oportunidade única para os cientistas sociais criarem novas bases de dados e desenvolverem investigações científicas originais (HOPKINS e KING, 2010). No entanto, essas informações, muitas vezes, não são disponibilizadas de uma maneira que facilite a coleta, a estruturação e a análise dos dados. Por exemplo, ao invés de serem ofertadas em planilhas de trabalho ou em bases de dados para *softwares* de análise estatística (SPSS, Stata, R, SAS, etc), as informações são veiculadas em páginas html, em formato de imagem ou documento portátil (.pdf), o que dificulta e/ou inviabiliza a coleta sistemática dos dados e, consequentemente, limita o avanço do conhecimento científico em diferentes áreas de pesquisa.

Embora amplamente empregadas na área da computação, existem poucas aplicações de coleta automatizada de dados nas Ciências Sociais[1]. A maioria dos estudiosos ainda coleta os dados de suas pesquisas utilizando algum procedimento manual repetitivo, seja fazendo isso por conta própria ou através da contratação de estudantes de graduação que trabalham como "coletadores" e "tabuladores". Para Hopkins e King (2010), "*given the infeasibility of much larger scale human-based coding, the need for automated methods is growing fast*" (HOPKINS e KING, 2010: 229). Os procedimentos automatizados de coleta reduzem o custo, o tempo e a probabilidade de erros que podem levar a inferências enviesadas, especialmente em pequenas amostras[2]. Em particular, em relação à confiabilidade dos dados, a coleta e a tabulação manual de informações podem gerar consequências catastróficas. Por exemplo, recentemente veio à tona o caso de Reinhart e Rogoff que supostamente cometeram um erro de cálculo em uma planilha de Excel em um trabalho amplamente citado, *Growth in a Time of Debt (2010)*[3], e que foi utilizado para orientar a política macroeconômica de diferentes países[4]. Em resumo, a adoção de instrumentos automáticos de extração de informações pode gerar base de dados mais baratas, mais rápidas e mais confiáveis.

Este artigo apresenta uma nova ferramenta para a coleta automatizada de dados: o TOGARY[5]. O objetivo do programa é extrair as informações das sentenças condenatórias por improbidade administrativa catalogadas pelo Cadastro Nacional de Condenações Cíveis por Ato de Improbidade Administrativa e Inelegibilidade[6], criado pelo Conselho Nacional de Justiça (CNJ). O sistema é aberto à consulta pública e disponibiliza alguns detalhes referentes às sentenças condenatórias por Improbidade Administrativa[7]. A Figura 1 ilustra a interface da ferramenta de busca criada pelo CNJ.

[1] Essas aplicações nas Ciências Sociais tendem a estarem restritas ao software R Statistical (BARBERA, 2013).

[2] Ver Munzert et al (2015) para uma boa introdução ao *webscraping* e de mineração de texto. Rei e Lowe (2003), Hopkins e King (2010), Grimer e King (2011), Grimer e Stewart (2013) e Orazio et al (2014) aplicam métodos automatizados, mas para análise de conteúdo.

[3] Ver: <http://galileo.stmarys-ca.edu/awilliam/Winter%202012-%20Moraga%20-%20Saturday%20and%20Santa%20Clara%20-%20GMAN%20503/documents/Growth_in_Time_Debt-reinhardandrogoff.pdf>

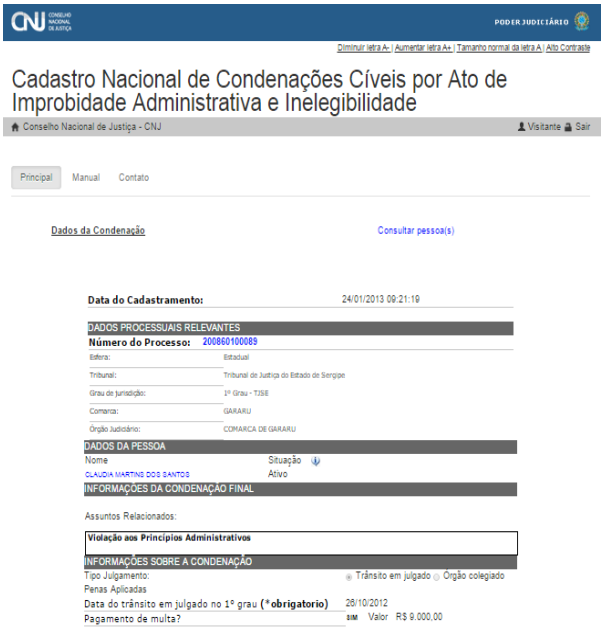
[4] Ver: <http://www.newyorker.com/news/john-cassidy/the-reinhart-and-rogoff-controversy-a-summing-up> e <http://www.publico.pt/economia/noticia/estudo-usado-como-argumento-para-a-austeridade-colocado-em-causa-1591613>.

[5] O nome do programa é uma homenagem ao professor Gary King, da Universidade de Harvard.

[6] Para mais detalhes, ver: <http://www.cnj.jus.br/improbidade_adm/consultar_requerido.php>.

[7] Para mais informações, ver: <http://www.planalto.gov.br/ccivil_03/leis/L8429.htm>.

Figura 1 – Interface Cadastro Nacional de Condenações Cíveis por Ato de Improbidade Administrativa e Inelegibilidade.



Fonte: CNJ (2015)

Nossa motivação para criar o *software* é a limitada transparência institucional uma vez que os dados originais apenas estão disponíveis para consulta caso-a-caso, o que inviabiliza a coleta manual. A atualização mais recente do TOGARY tem informações para mais de 15.000 casos de improbidade e outros delitos julgados pelas esferas estadual, federal, eleitoral e superior no Brasil entre 1992 e 2014. Em termos substantivos, esse trabalho faz três principais contribuições à literatura. Em primeiro lugar, utiliza e disponibiliza publicamente uma base de dados original com informações observacionais ao invés de examinar dados de percepção. Segundo, analisamos dados desagregados por seções judiciais, cidades, unidades da federação e Tribunais Regionais Federais (TRFs) ao invés de trabalhar com informações agregadas por país. Por fim, estimamos um indicador de corrupção a partir do tempo de tramitação das sentenças condenatórias ao invés de focar na incidência do fenômeno. Nossa proposta metodológica reduz eventuais erros sistemáticos e aleatórios de mensuração que são especialmente recorrentes na criação de indicadores subjetivos. Além disso, a utilização de dados agregados por país limita a capacidade de explicações no nível individual e, dessa forma, deve-se ter bastante cuidado com problemas de falácia ecológica (KING, 1997). Com este trabalho, esperamos difundir a utilização de procedimentos automatizados de coleta de dados em pesquisas empíricas em Ciência Política. Além disso, esperamos expandir a utilização da ciência para combater a corrupção no Brasil.

O restante do artigo está organizado da seguinte forma. A próxima seção apresenta uma breve revisão da literatura sobre corrupção. Depois disso, apresentamos um *overview* sobre a metodologia *webscraping* e como o TOGARY funciona e deve ser utilizado. A seção seguinte descreve brevemente o funcionamento do sistema judicial no Brasil. Para ilustrar a funcionalidade do programa, utilizamos estatística descritiva e análise espacial na seção dos resultados. A última parte sumariza as conclusões.

REVISÃO DA LITERATURA SOBRE CORRUPÇÃO[8]

O estudo da corrupção[9], não só tem importância empírica, mas também é metodologicamente desafiador (KNACK, 2006). Por exemplo, Chang (2005) afirma que "*political corruption is considered one of the most destructive yet unresolved problems common to most countries*" (CHANG, 2005: 717). Como qualquer atividade criminosa, as práticas locumpletativas tendem a ser subestimadas, uma vez que os agentes têm incentivos para minimizar a visibilidade de suas ações. Por um lado, a crescente oferta de banco de dados transversais e longitudinais permitiram grande avanço nas pesquisas sobre as causas e consequências da corrupção (CHANG, 2005; TREISMAN, 2007). Para Gelhbach (2009), os estudos comparados sobre o tema utilizam duas principais fontes de dados: (1) questionários com empresas e indivíduos e (2) *surveys* com especialistas. Os indicadores mais conhecidos são os da Transparência Internacional[10] e do Banco Mundial[11]. Por outro lado, a maior parte da literatura utiliza dados agregados de percepção que são naturalmente subjetivos e, dessa forma, mais sujeitos a erros de mensuração. De acordo com Jong-Sung e Khagram (2005), estudos estatísticos sobre corrupção são dificultados pela falta de dados quantitativos confiáveis (JONG-SUNG e KHAGRAM, 2005:136). Além disso, a utilização de dados agregados por país limita a capacidade de inferências válidas no nível individual (ROBINSON, 1951; KING, 1997). Diante dessas limitações, as inferências baseadas nessas informações devem ser tratadas com cautela[12].

A corrupção é uma característica inerente às sociedades humanas no espaço e no tempo (AIDT, 2003). Comparativamente, a literatura empírica foca em três dimensões explicativas fundamentais: (1) econômica, (2) política, e (3) cultural (YOU e KHAGRAM, 2005). Por exemplo, no que se refere às questões econômicas, há evidências de que a corrupção está correlacionada com a desigualdade de renda (YOU e KHAGRAM, 2005) e possui um efeito negativo sobre o crescimento econômico (MAURO, 1995). Treisman (2000) também encontra correlações significativas entre níveis de desenvolvimento econômico e medidas agregadas de percepção de corrupção. Em particular, depois de controlar por diferentes variáveis, Treisman (2000) reporta um coeficiente de aproximadamente -4. Braun e Di Tella (2004) argumentam que a alta e a variação das taxas de inflação fazem com que haja um aumento nos contratos referentes gastos do governo, tornando difícil a fiscalização dos contratos de referentes à prestação de serviço e fazendo com que a corrupção aumente. Ades e Di Tella (1999) argumentam que a corrupção tende a ser maior em locais com alta renda econômica, o que facilitaria tal prática. Além disso, há também uma correlação sistemática entre recursos naturais e os níveis de corrupção (ADES e DI TELLA, 1999; GYLFASON, 2001; LEITE e WEIDMAN, 1999).

Quanto às explicações políticas da corrupção, os estudiosos costumam utilizar democracia (MONTINOLA e JACKMAN, 2002), o tamanho do governo (LAPALOMBARA, 1994; FRIEDMAN et al., 2000; LA PORTA et al., 1999), a descentralização (TREISMAN, 2000; FISMAN e GATTI, 2002), a liberdade de imprensa (BRUNETTI e WEDER, 2003; TREISMAN, 2007), entre outros argumentos institucionais como variáveis independentes para a explicação do fenômeno. Panizza (2001) argumenta que os sistemas presidenciais estão associados significativamente a

[8] A literatura nacional empírica sobre as causas e consequências da corrupção ainda é bastante limitada. Importantes exceções podem ser encontradas em: Albuquerque e Ramos (2006), Ferraz e Finan (2007), Ferraz, Finan e Moreira (2008), Pereira, Melo e Figueiredo (2008), Filgueiras (2009), Leite (2010), Henrique e Ramos (2011) e Batista (2013). Ainda, é possível citar os trabalhos do professor Bruno Speck.

[9] Entenda-se corrupção como o uso indevido do cargo público para benefício privado. O que também pode ser tratado como improbidade administrativa, de acordo com a Lei nº 8.249, de 2 de junho de 1992. Em 2013, a Lei 12.846 de 1 de agosto, dispôs sobre a responsabilização administrativa e civil de pessoas jurídicas pela prática de atos contra a administração pública, nacional ou estrangeira. Ver <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2013/lei/l12846.htm>. Lancaster e Montinola (1997), em *Toward a Methodology for the Comparative Study of Political Corruption*, apresentam uma revisão de várias definições conceituais.

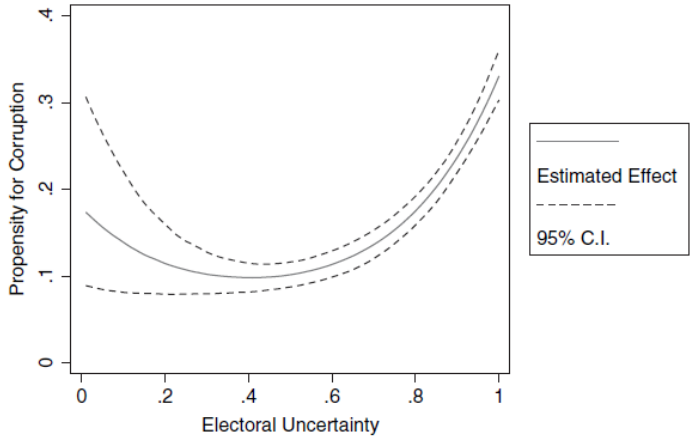
[10] Ver < https://www.transparency.org/>

[11] Ver < http://data.worldbank.org/indicator/IQ.CPA.TRAN.XQ>

[12] Diferente da maior parte da literatura que trabalha com informações agregadas por país, Chang (2005) utiliza um banco de dados com casos individuais (por legislador) de corrupção na Itália.

maiores percepções sobre corrupção. Kunicová e Rose-Ackerman (2005) observam que a corrupção aumenta quando presidencialismo é combinado com lista fechada de representação proporcional (RP), em eleições legislativas. Por um lado, Golden e Chang (2007) indicam uma correlação positiva entre corrupção e magnitude do distrito em sistemas de RP de lista aberta. Por outro, Persson, Tabellini e Trebbi (2003) reportam que "*corruption is also higher in countries electing fewer candidates per district*" (PERSSON, TABELLINI e TREBBI: 960). Nyblade e Reed (2008), ao analisarem a relação entre competição eleitoral e corrupção no Japão, distinguem dois tipos de atividades corruptas: *looting* e *cheating*. A primeira diz respeito a ganhos materiais enquanto a segunda refere-se a atividades ilegais com propósitos eleitorais, como a compra de voto, por exemplo. Chang (2005) relata uma associação não linear entre competição política (*electoral uncertanntiy*) e níveis de corrupção. A Figura 2 ilustra os resultados reportados por Chang (2005) para o caso da Itália.

Figura 2 - Relação estimada entre incerteza eleitoral e corrupção: o paradoxo da instabilidade



Fonte: Chang (2005)

O eixo X representa a probabilidade de corrupção e o eixo Y representa o nível de incerteza eleitoral (competição política). Como pode ser observado, a corrupção é maior nos dos extremos, ou seja, quando há pouca competição e quando existe muita incerteza eleitoral. Rose-Ackerman (1999) define esse fenômeno como o paradoxo da estabilidade. Para explicar, quando os políticos estão muito confiantes que serão reeleitos, aumenta-se os incentivos para adoção de práticas ilegais já que a chance de punição eleitoral é pequena. Por outro lado, quando a competição é muito intensa, os custos eleitorais tendem a aumentar, o que incentiva os políticos a arrecadarem mais doações de campanha. No original, "*while too much uncertainty of winning reelection encourages a legislator to be corrupt, too much uncertainty of victory can drive him, too*" (CHANG, 2005: 726)[13].

Por fim, as explicações culturais investigam os efeitos da religião e do fracionamento étnico (Mauro, 1995), a experiência colonial (TREISMAN, 2000), o sistema legal (LA PORTA ET AL., 1999; TREISMAN, 2000; SANDHOLTZ e KOETZLE, 2000; PALDAM, 2001) e gênero (DOLLAR, FISMAN e GATTI, 2001; SWAMY et al, 2001) sobre a corrupção. A maioria das evidências é baseada em medidas de percepção, do que por definições substantivas. Portanto, a literatura sobre a corrupção ainda enfrenta muitos desafios metodológicos. O obstáculo mais complicado é encontrar um critério válido para mensurar corrupção. A falta de uma medida confiável reduz a capacidade de estudiosos se envolver em projetos empíricos comparativos (TREISMAN, 2000).

[13] A noção de que a competição política pode afetar os níveis de corrupção também pode ser encontrada nos Federalistas (nº 72).

ACERCA DO WEBSCRAPING

O acesso à informação nos meios eletrônicos, em alguns casos, é um grande problema para pesquisadores e estudiosos. Muitas vezes, quando disponíveis, as informações estão em formatos inoperáveis em programas de análise de dados. Além disso, atualmente, "*quite a lot of researchers are working on extracting information about types of events, entities or relationships from textual data*" (VARGIU e URRU, 2012).

A lógica do *webscraping* é criar um programa (*bot*) que simula as ações de usuários físicos (POGGI et al, 2007) e seleciona as informações relevantes para o pesquisador dentro da página web (FERNÁNDEZ-VILLAMOR et al, 2011), para que, a partir disso, os dados possam ser analisados. Vargiu e Urru (2012) também apontam que o foco dessa atividade é transformar dados despadronizados em base de dados estruturadas. Por fim, Pan et al (2002) identificam que as duas principais tarefas desse procedimento são acessar as páginas que contêm as informações desejadas e obter a estrutura das informações dentro das páginas HTML. Essa técnica é comumente utilizada para conversão monetária no comércio de câmbio, no monitoramento de informações climáticas, na detecção de mudanças em websites, em pesquisas web e integração de dados, além da análise de *merchandising* (VARGIU e URRU, 2012).

Penman, Baldwin e Martinez (2009) apontam que uma das principais dificuldades para esse método de coleta são as camadas de estilo presentes nas páginas que podem mascarar as informações dentro da página HTML[14]. Os autores, complementarmente, afirmam que: "*these styles change over time as each website is updated with additional content or a new layout. This makes working with data across websites cumbersome*" (PENMAN, BALDWIN e MARTINEZ, 2009).

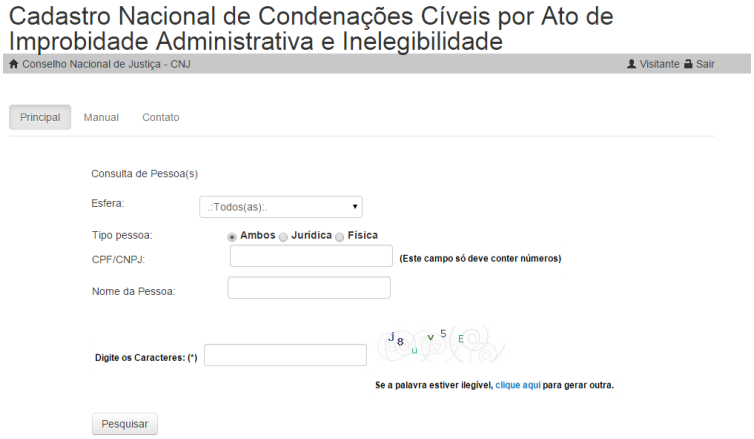
Seguindo essa mesma lógica, Vargiu e Urru (2012) apontam que alguns sistemas *web* impõe algumas barreiras que impedem a extração automática. Um exemplo de entrave é a mudança dos nomes das classes e atributos nas páginas html, que faz com que o programa perca a indexação dos parâmetros de busca (HARTMANN, COLLINS e KLEMMER, 2007). Além disso, há a questão do uso de cookies, autenticação e protocolos de segurança que dificultam o acesso aos servidores web (PAN et al, 2002). Em alguns casos, resta apenas a coleta manual como mecanismo que obtenção das informações, porém Vargiu e Urru (2012) destacam que "*although sometimes this is the only way to export information from a Web page, this is not feasible in practice, especially for big company projects, being too expensive*".

O que o TOGARY faz e como funciona?

O CNJ disponibiliza uma base original sobre as sentenças de improbidade administrativa julgadas por todas as esferas judiciais (estadual, federal, superior e eleitoral). O sistema permite uma busca a partir de determinados parâmetros: (1) esfera – estadual, federal, superior e eleitoral; (2) localidade – todos os estados da federação ou as regiões federais; (3) tipo de pessoa – indivíduo ou organização; (4) CPF/CNPJ – número único que identifica a pessoa perante à Receita Federal (RFB); (5) nome da pessoa – descrição do indivíduo ou da organização. A Figura 3 mostra como os parâmetros de busca estão dispostos.

[14] Abreviação de *HyperText Markup Language*, que uma linguagem de diagramação das informações utilizada para a produção de páginas web.

Figura 3 – Página inicial do Cadastro de Improbidade Administrativa



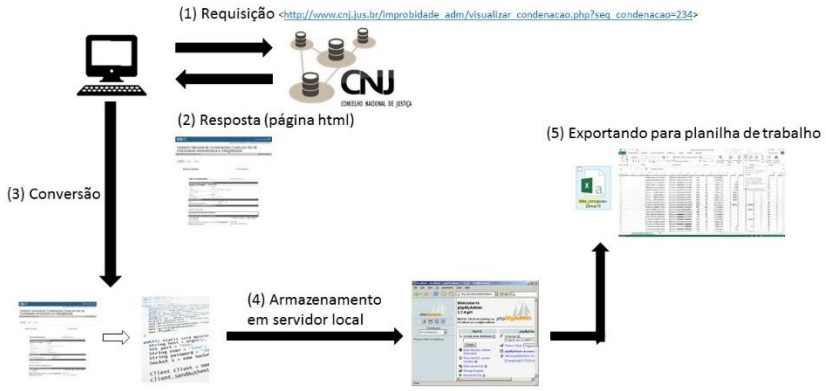
Fonte: CNJ (2015)

Ainda dentro do sistema do CNJ é possível observar quando a ação foi proposta bem como a sua data de julgamento. Dessa forma, tem-se um indicador de morosidade judicial. Outro dado interessante diz respeito as penas aplicadas em cada sentença, o que permite criar um indicador de severidade judicial. No entanto, em função da indisponibilidade da coleta sistemática no sistema do CNJ, já que as informações somente estão disponibilizadas caso-a-caso, o pesquisador acaba restringindo-se à coleta manual, que demanda muito tempo e recurso e possui uma alta probabilidade de erros ao longo do processo de obtenção dos dados. Além disso, o sistema do CNJ é atualizado periodicamente, o que pode dificultar, a longo prazo, a consolidação de uma base representativa às informações presentes no CNJ. Ou seja, tem-se um "tesouro informacional" mas que não está publicamente disponível para pesquisadores e cidadãos.

O TOGARY foi projetado a fim de superar essas limitações. Do ponto de vista técnico, o programa foi desenvolvido na linguagem de programação Java, por meio da biblioteca Jsoup[15]. Inicialmente, o programa requisita as páginas das condenações dos servidores do CNJ por meio de uma URL[16] informando o número da sequência de condenação[17](1). Após isso, é retornada uma página (a nível de representação, ver Figura 1; para mais especificações, ver Código 1) contendo as informações referentes a sentença (2), onde elas são convertidas em uma única *string* (3), o Código 2 descreve esse procedimento aplicado na linguagem de programação. A partir dessa *string*, as informações desejadas são extraídas e armazenadas em um servidor local (4). Terminado a etapa de coleta das condenações, o mesmo processo é aplicado para a obtenção das informações referentes aos processos, já que o CNJ as disponibiliza em bases distintas e que o programa realiza o cruzamento entre elas. A única diferença entre esses dois procedimentos é a URL do processo[18] que difere do padrão das condenações. Após esses procedimentos, nós exportamos todas as informações em formato de planilha, na extensão de valores separados por vírgulas (.csv). A Figura 4 ilustra o funcionamento do programa.

[15] Mais informações, ver: <http://jsoup.org/>.
[16] Por exemplo: <www.cnj.jus.br/improbidade_adm/visualizar_condenacao.php?seq_condenacao=234>.
[17] Esse número é inteiro, positivo e maior que 0. Além disso, como a base de dados do CNJ é atualizada periodicamente, seu valor varia sempre. Até o momento da elaboração do artigo, constavam 24.436 sentenças na base. O programa gera os números aleatoriamente dentro do intervalo.
[18] Ver: <http://www.cnj.jus.br/improbidade_adm/visualizar_processo.php?seq_processo=4947>.

Figura 4 - Diagrama de funcionamento do TOGARY



Fonte: elaboração dos autores (2015)

Código 1 – Trecho do código da página de resposta do CNJ

```
[...]
<table width="700px" align="center" border="0" style='margin-left: 125px; _margin-
left: 0px;'>
   <tr>
      <td class='td_form'  nowrap width=20%>
        <span ><b>Data do Cadastramento:</b></span>
      </td>
      <td >
                          07/07/2010 13:29:29
      </td>
   </tr>
   <tr>
      <td align="left" colspan="2">
                                     
      </td>
   </tr>
    <tr bgcolor="#666666">
      <td align="left" colspan="2">
        <span class="label2">DADOS PROCESSUAIS RELEVANTES</span>
      </td>
   </tr>
  <tr>
       <td colspan="2">
         <table width="700px" border="0">
             <tr>
               <td class='td_form' nowrap="nowrap" width="27%">
                 <span ><b>Número do Processo:</b></span>
               </td>
               <td width="85%"  style="font-color: blue"><b><a
               href="visualizar_processo.php?seq_processo=1767">
                <font color='blue'>00663106520068220009</font></a> </b>
               </td>
             </tr>
             <tr>
                <td colspan='2'>
                  <div class='areaHierarquia'>
                       <div id='hierarquia'>
                       <div id='hierarquia-linha-esfera' class='hierarquia-orgaos-
linha'>
                       <div                class='hierarquia-orgaos-coluna-esquerda'
id='hierarquia-coluna-esquerda-esfera'>

[...]
```

Código 2 - Método de captura das informações do processos e das condenações

```
public String capturar_processo(int seq_condenacao) throws IOException {
//início do método
```

```
        String url =
"http://www.cnj.jus.br/improbidade_adm/visualizar_condenacao.php?seq_condenacao="
+ seq_condenacao;
        doc = Jsoup.connect(url).get();
        if (verificaPermissao(doc.toString())) {

            List < String > processos = new ArrayList < String > ();
            Elements tables = doc.select("table");
            String informacoes = null;
            int j;
            int i = j = 0;

            Element table = tables.get(5);

            for (Element row: table.select("tr")) {
                j = 0;
                for (Element column: table.select("td")) {
                    if (i == 0 && j == 0) {
                        informacoes = column.text().toString();
                    }
                    j++;
                }
                i++;
            }
            System.out.println(informacoes);
            return informacoes;
        }

        return "PROCESSO INACESSÍVEL!";

    } //fim do método
```

Por fim, realizamos *data cleaning* das informações originais e os dados são salvos e disponibilizados nos seguintes formatos: .xls, .sav e .dta. A atualização mais recente TOGARY tem informações sobre mais de 15.000 casos de corrupção julgados entre 1992 e 2014. Todas as informações estão publicamente disponíveis em: <http://dx.doi.org/10.7910/DVN/27787>.

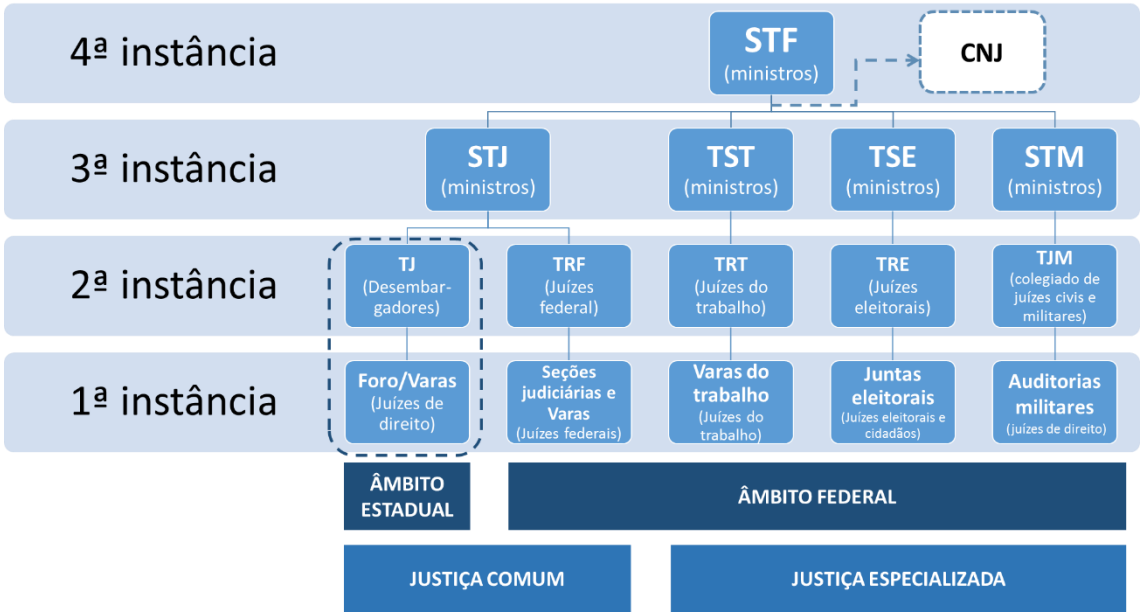
ACERCA DO SISTEMA JUDICIÁRIO BRASILEIRO

De acordo com o artigo 92, da Constituição Federal de 1988, o sistema judiciário brasileiro é composto pelos seguintes órgãos: Supremo Tribunal Federal (STF), Conselho Nacional de Justiça (CNJ), Supremo Tribunal de Justiça (STJ), Tribunais Regionais Federais (TRFs) e seus juízes, Tribunal Superior do Trabalho (TST), Tribunal Superior Eleitoral (TSE), Tribunal Superior Militar (TJM) – no âmbito da União – e Tribunais Estaduais e do Distrito Federal e seus juízes – no âmbito federativo, onde todos eles gozam de autonomia administrativa e financeira (BRASIL, 1998). Os Tribunais de Justiça Estaduais e do Distrito federal e os Federais fazem parte da justiça comum, enquanto os Tribunais Eleitoral, Militar e do Trabalho fazem parte da justiça especializada.

Inicialmente, os processos são julgados em primeira instância por juízes. Cabendo recurso, são julgados em segunda instância por desembargadores no âmbito federativo e por juízes regionais no âmbito da União. Caso se estendam, são analisados por ministros dos Tribunais Superiores e, por fim, envolvendo litígios referentes a questões constitucionais, o STF é o responsável pelo veredito final nos processos, conforme a Figura 3.

Em casos especiais, alguns processos podem ser iniciados em instâncias superiores. Por exemplo, em casos de infrações penais ou crimes comuns, parlamentares federais, presidentes e ministros do executivo têm a prerrogativa de serem julgados pelo STF. Os governadores e deputados estaduais e distritais são julgados pelo STJ. Enquanto os prefeitos são julgados na segunda instância da justiça comum.

Figura 5 - Organograma do Poder Judiciário no Brasil



Fonte: elaborado pelos autores (2015) com base nas informações presentes no capítulo III da CF (BRASIL, 2014)

Justiça Federal

Supremo Tribunal Federal

Órgão máximo da justiça brasileira, o STF, além de última instância de julgamento, é responsável por examinar: ações diretas de inconstitucionalidade, litígios entre algum ente federativo com um Estado ou organismo internacional, extradição de qualquer cidadão estrangeiro (BRASIL, 1998). É composto por onze ministros, indicados pelo Presidente da república e analisado pelo Senado Federal.

Comum

A justiça comum federal é composta pelas varas, seções judiciárias, tribunais regionais e juízes federais. Ela apresentada na seção IV, no capítulo III, da CF. Sua função é julgar crimes e infrações contra a União.

Em relação aos tribunais regionais, eles são organizados em cinco regiões, primeira região é composta pelos estados do Acre, Amazonas, Amapá, Bahia, Distrito Federal, Goiás, Maranhão, Minas Gerais, Mato Grosso, Pará, Piauí, Rondônia, Roraima e Tocantins. A segunda por Rio de Janeiro e Espírito Santo. A terceira por São Paulo e Mato Grosso do Sul. A quarta por Rio Grande do Sul, Santa Catarina e Paraná. E, por fim, a quinta região por Alagoas, Ceará, Paraíba, Pernambuco, Rio Grande do Norte e Pernambuco, de acordo com o Mapa 1.

Mapa 1 - Distribuição dos TRFs no país

Fonte: disponível em: <http://www.jf.jus.br/>.

Trabalho

A Justiça federal do trabalho é do ramo da justiça especializada, conforme a seção V, do capítulo III. Ela é a responsável pela mediação dos litígios laborais entre patrões e funcionários. Suas competências são definidas na Emenda Constitucional nº 45, de 2004. É composta por juízes trabalhistas nas varas e nos tribunais regionais do trabalho e por ministros no TST

Eleitoral

A justiça eleitoral é responsável pela definição das regras eleitorais e pela organização, controle e apuração das eleições. Tem o poder de decretar a perda de mandatos e a inelegibilidade de candidatos envolvidos em irregularidades. É composta por juízes nas juntas e tribunais regionais e por ministros no TSE.

Militar

A justiça militar é o ramo da justiça especializada responsável por julgar crimes cometidos por militares, de qualquer natureza, definidos em lei. É composta por juízes civis nas auditorias, por juízes civis e militares nos tribunais de justiça e por ministros no STM.

Justiça Estadual

Comum

É composta por juízes de direito que atuam em fórum e varas e por desembargadores nos tribunais de justiça. Sua função é julgar qualquer causa que não seja da competência dos órgãos especializados. Ela concentra o maior número de litígios no Brasil (STF, 2011)

CNJ

Criado a partir da EC nº 45, de 2004, o Conselho Nacional de Justiça é o órgão responsável pelo controle administrativo e financeiro do Poder Judiciário. É composto por quinze membros, com mandato de dois anos, das seguintes áreas do judiciário brasileiro: o presidente do STF, um ministro do STJ, um ministro do TST, um desembargador do TJ, um juiz estadual, um juiz do TRF, um juiz federal, um juiz do TRT, um juiz do trabalho, um membro do Ministério Público da União (MPU), um membro do Ministério Público Estadual (MPE), dois advogados indicados pela Ordem dos Advogados do Brasil (OAB) e dois cidadãos de notório saber indicados pela Câmara dos Deputados e pelo Senado. O CNJ é presidido pelo presidente do STF.

RESULTADOS

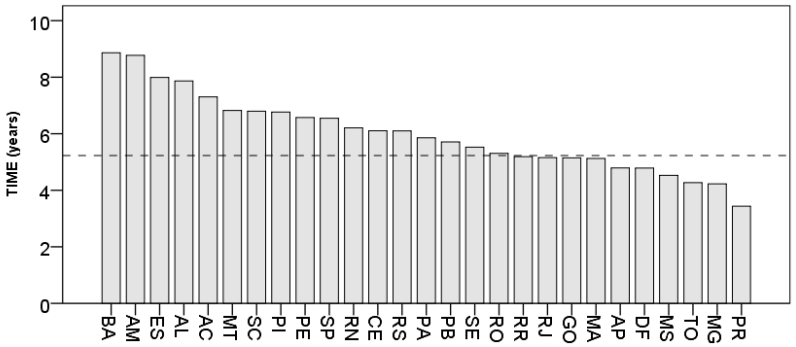
Em média, a justiça brasileira leva 5,23 anos para julgar os casos de corrupção, com um desvio padrão de 3,38 anos, de acordo com a Tabela 1. Bahia (8,86), Amazonas (8,77), Espírito Santo (7,99) e Alagoas (7,86) são os estados que levam mais tempo para julgar, enquanto Amapá (4,79), Mato Grosso do Sul (4,53), Tocantins (4,27), Minas Gerais (4,23) e Paraná (3,44) são os estados que julgam mais rápido, conforme o Gráfico 1 e o Mapa 1.

Tabela 1 – Tempo médio de julgamento dos casos de corrupção (anos)

N	Min	Max	Média	Desvio padrão
15.358	0,08	22,48	5,23	3,38

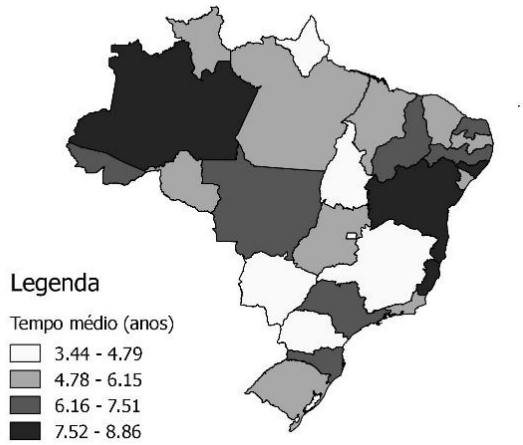
Fonte: elaborado pelos autores (2015)

Gráfico 1 – Tempo médio de julgamento dos casos de corrupção por estado (anos)



Fonte: elaborado pelos autores (2015)

Mapa 1 – Tempo médio de julgamento dos casos de corrupção por estado (anos)



Fonte: elaborado pelos autores (2015)

Comparativamente, a partir da Tabela 2, pode-se constatar que não há diferença no tempo de julgamento dos casos de corrupção entre as esferas estaduais e federais. Contudo, o tempo de julgamento da esfera federal é mais homogêneo que a estadual, a partir do coeficiente de variação.

Tabela 2 – Tempo médio de julgamento dos casos de corrupção por esfera judicial (anos)

Esfera	N	Média	Desvio padrão	Coeficiente de variação
Estadual	11.959	5,22	3,48	0,67
Federal	2.399	5,32	2,86	0,54

t = -1,486; p-value = 0,137

Fonte: elaborado pelos autores (2015)

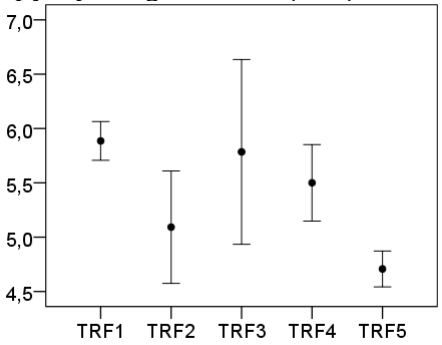
Analisando a esfera federal, a análise de variância (ANOVA) indica que há uma diferença significativa entre o tempo de julgamento dos Tribunais Regionais Federais (TRFs). A 5ª região (TRF5) possui o menor tempo de julgamento (4,71 anos), enquanto a 1ª região (TRF1) possui o maior tempo (5,89 anos), a Tabela 3 e o Gráfico 2 sumarizam essas informações sobre os TRFs.

Tabela 3 – Tempo médio de julgamento dos casos de corrupção por região federal (anos)

TRF	N	Média	DP	CV
1	920	5,89	2,76	0,47
2	173	5,09	3,45	0,68
3	71	5,78	3,59	0,62
4	292	5,50	3,05	0,55
5	943	4,71	2,58	0,55

F = 21,563; p-value <0,000

Gráfico 2 – Tempo médio de julgamento dos casos de corrupção por região federal (anos)

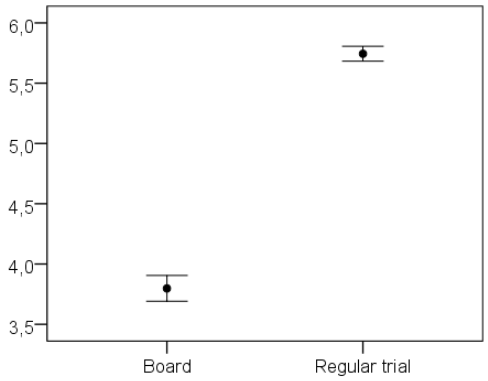


Fonte: elaborado pelos autores (2015)

As decisões tomadas por órgãos colegiados (3,8 anos) são mais rápidas que as tomadas por trânsito julgado (5,74 anos). Como os interceptos das médias não se cruzam, pode-se afirmar que as diferenças

médias são estatisticamente significativas (t = 30,910; p-value<0,000), conforme exposto no Gráfico 3.

Gráfico 3 – Tempo médio de julgamento por tipo (anos)



Fonte: elaborado pelos autores (2015)

Após a criação do CNJ, a partir de 2004, houve uma diminuição no tempo de julgamento dos casos de corrupção, conforme a Tabela 4. Controlamos também essa diferença por esfera (Tabela 5 e Gráfico 4), uf (Tabela 6), TRFs (Tabela 7 e Gráfico 6) e tipo de julgamento (Tabela X), onde todas elas são estatisticamente significativa.

Tabela 4 – Diferença no tempo médio de julgamento dos casos de corrupção antes e após a criação do CNJ (anos)

Criação CNJ	N	Média	Desvio padrão	Coeficiente de variação
Antes do CNJ	5.077	7,59	3,33	0,43
Após o CNJ	9.281	3,17	2,19	0,69

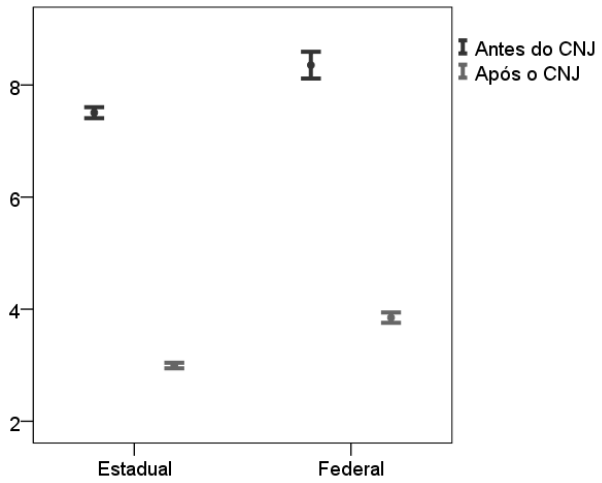
t = 85,103; p-value = 0,000
Fonte: elaborado pelos autores (2015)

Tabela 5 – Diferença no tempo médio de julgamento dos casos de corrupção antes e após a criação do CNJ por esfera judicial (anos)

Esfera	Criação CNJ	N	Média	Desvio padrão	Coeficiente de variação
Estadual	Antes do CNJ	4.567	7,51	3,38	0,45
	Após o CNJ	7.392	2,99	2,19	0,73
Federal	Antes do CNJ	510	8,35	2,74	0,33
	Após o CNJ	1.889	3,85	2,04	0,53

Estadual: t = 80,396; p-value = 0,000
Federal: t = 34,577; p-value = 0,000
Fonte: elaborado pelos autores (2015)

Gráfico 4 - Diferença no tempo médio de julgamento dos casos de corrupção antes e após a criação do CNJ por esfera judicial (anos)



Fonte: elaborado pelos autores (2015)

Tabela 6 – Diferença no tempo de julgamento dos casos de corrupção antes e após a criação do CNJ por unidade da federação (anos)

UF	Criação CNJ	N	Média	Diferença média	P-valor	Desvio padrão	Coeficiente de variação
AC	Antes do CNJ	20	9,2	4,48	0,000	2,526	0,275
	Após o CNJ	25	4,72			2,301	0,488
AL	Antes do CNJ	10	10,8	6,01	0,000	1,476	0,137
	Após o CNJ	14	4,79			1,672	0,349
AM	Antes do CNJ	3	10,33	2,9	0,019	0,577	0,056
	Após o CNJ	7	7,43			1,618	0,218
AP	Antes do CNJ	17	8,94	5,91	0,000	2,436	0,272
	Após o CNJ	63	3,03			1,636	0,540
BA	Antes do CNJ	18	10,06	5,81	0,000	2,182	0,217
	Após o CNJ	8	4,25			2,252	0,530
CE	Antes do CNJ	11	8,91	4,67	0,000	2,343	0,263
	Após o CNJ	29	4,24			2,262	0,533
DF	Antes do CNJ	19	8,37	5,28	0,000	3,89	0,465
	Após o CNJ	68	3,09			2,91	0,942
ES	Antes do CNJ	94	11,55	7,85	0,000	3,222	0,279
	Após o CNJ	98	3,7			2,047	0,553
GO	Antes do CNJ	222	6,45	3,43	0,000	3,946	0,612
	Após o CNJ	240	3,02			2,321	0,769
MA	Antes do CNJ	54	8,04	4,54	0,000	2,767	0,344
	Após o CNJ	169	3,5			1,946	0,556
MG	Antes do CNJ	485	6,83	4,56	0,000	3,531	0,517
	Após o CNJ	1.027	2,27			2,073	0,913
MS	Antes do CNJ	17	7,65	4,5	0,000	3,316	0,433

	Após o CNJ	62	3,15			1,687	0,536
MT	Antes do CNJ	68	7,74	2,97	0,000	3,514	0,454
	Após o CNJ	61	4,77			2,807	0,588
PA	Antes do CNJ	11	7,55	2,92	0,000	0,688	0,091
	Após o CNJ	35	4,63			2,533	0,547
PB	Antes do CNJ	27	6,78	2,7	0,000	3,042	0,449
	Após o CNJ	39	4,08			2,095	0,513
PE	Antes do CNJ	12	9,83	6,2	0,000	1,642	0,167
	Após o CNJ	19	3,63			1,739	0,479
PI	Antes do CNJ	14	7,79	3,9	0,002	3,215	0,413
	Após o CNJ	9	3,89			1,9	0,488
PR	Antes do CNJ	466	7,44	5,28	0,000	3,565	0,479
	Após o CNJ	2.668	2,16			1,871	0,866
RJ	Antes do CNJ	40	5,98	1,7	0,000	2,547	0,426
	Após o CNJ	142	4,28			2,029	0,474
RN	Antes do CNJ	74	6,91	2,1	0,000	2,746	0,397
	Após o CNJ	100	4,81			1,674	0,348
RO	Antes do CNJ	363	6,26	3,12	0,000	3,336	0,533
	Após o CNJ	311	3,14			2,28	0,726
RR	Antes do CNJ	6	6,33	3,83	0,011	1,751	0,277
	Após o CNJ	4	2,5			1,915	0,766
RS	Antes do CNJ	445	6,99	3,12	0,000	3,457	0,495
	Após o CNJ	363	3,87			2,058	0,532
SC	Antes do CNJ	329	8,45	4,45	0,000	3,851	0,456
	Após o CNJ	308	4			2,189	0,547
SE	Antes do CNJ	17	9,47	5,88	0,000	1,586	0,167
	Após o CNJ	54	3,59			2,005	0,558
SP	Antes do CNJ	1.723	7,76	3,85	0,000	2,906	0,374
	Após o CNJ	1.399	3,91			2,036	0,521
TO	Antes do CNJ	2	7,5	3,71	0,001	0,707	0,094
	Após o CNJ	70	3,79			1,463	0,386

Fonte: elaborado pelos autores (2015)

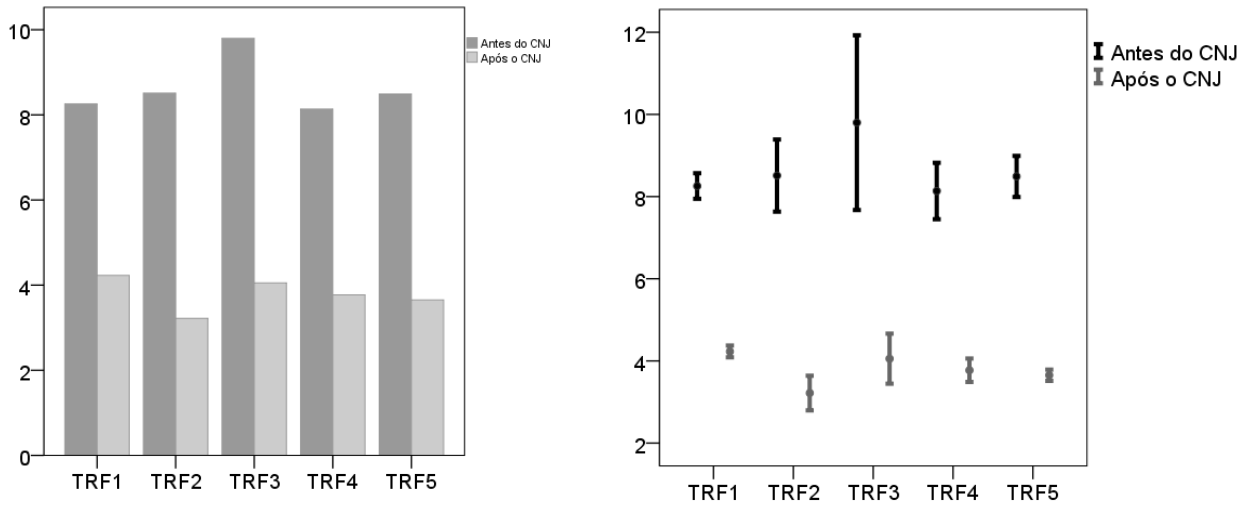
Tabela 6 – Diferença no tempo de julgamento dos casos de corrupção antes e após a criação do CNJ por TRF (anos)

TRF	Criação CNJ	N	Média	Diferença média	P-valor	Desvio padrão	Coeficiente de variação
1	Antes do CNJ	261	8,26	4,03	0,000	2,56	0,31
	Após o CNJ	659	4,23			1,89	0,45
2	Antes do CNJ	45	8,51	5,29	0,000	2,92	0,34
	Após o CNJ	128	3,22			2,41	0,75
3	Antes do CNJ	15	9,80	5,75	0,019	3,84	0,39
	Após o CNJ	56	4,05			2,28	0,56
4	Antes do CNJ	81	8,14	4,37	0,000	3,10	0,38

5	Após o CNJ	211	3,77	4,84	0,000	2,11	0,56
	Antes do CNJ	108	8,49			2,62	0,31
	Após o CNJ	835	3,65			2,02	0,55

Fonte: elaborado pelos autores (2015)

Gráfico 5 – Diferença no tempo médio de julgamento dos casos de corrupção antes e após a criação do CNJ na esfera federal (anos)



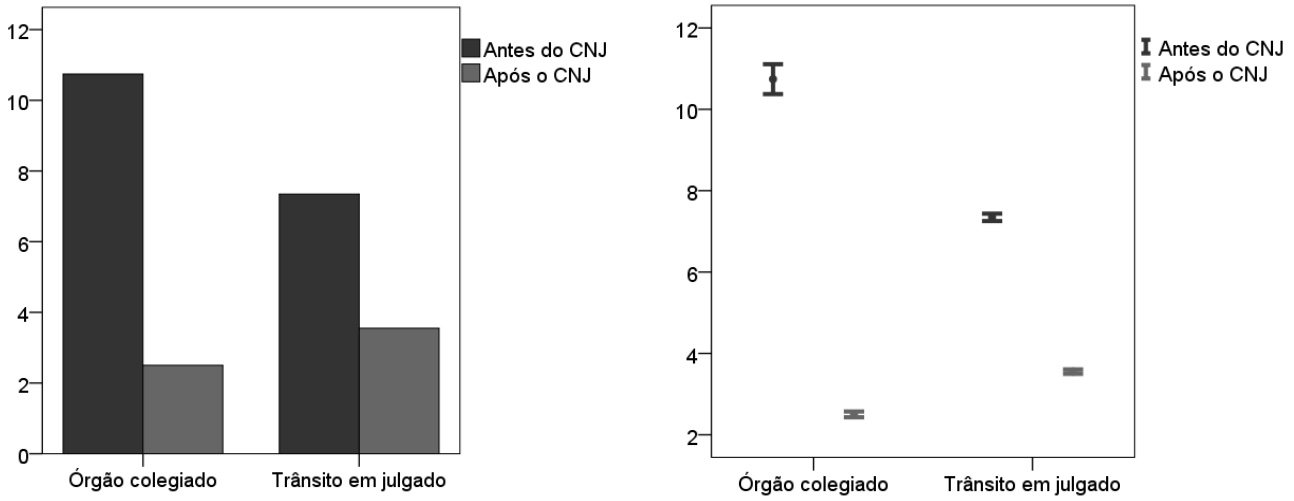
Fonte: elaborado pelos autores (2015)

Tabela 7 – Diferença no tempo médio de julgamento dos casos de corrupção antes e após a criação do CNJ por tipo de julgamento (anos)

Esfera	Criação CNJ	N	Média	Desvio padrão	Coeficiente de variação
Órgão colegiado	Antes do CNJ	368	10,74	3,60	0,34
	Após o CNJ	3.402	2,50	2,11	0,84
Trânsito em julgado	Antes do CNJ	4.709	7,35	3,18	0,43
	Após o CNJ	5.879	3,55	2,14	0,60

Órgão colegiado: t = 43,130; p-value = 0,000
Trânsito em julgado: t = 70,092; p-value = 0,000
Fonte: elaborado pelos autores (2015)

Gráfico 6 – Diferença no tempo médio de julgamento dos casos de corrupção antes e após a criação do CNJ por tipo de julgamento (anos)



Fonte: elaborado pelos autores (2015)

CONCLUSÕES

Atualmente, mensurar a corrupção de uma maneira adequada é um dos procedimentos mais difíceis enfrentados pelos pesquisadores e formuladores de políticas. Isso porque os agentes corruptos têm incentivos diretos para mascarar suas ações. Por esta razão, a maioria das produções científicas sobre as causas e as consequências da corrupção é baseada em percepção, medida por meio da realização de *surveys*. Apesar de serem muito importantes, os achados empíricos de *surveys* não são muito confiáveis, já que as respostas das perguntas podem ser influenciadas por fatores externos, tais como meios de comunicação, redação, abertos contra perguntas fechadas, etc.

No Brasil, o Conselho Nacional de Justiça fez uma importante contribuição ao estudo da corrupção observada quando disponibilizou os dados relativos aos processos de improbidade administrativa para todas as esferas judiciais (subnacional, nacional, eleitoral e superior). Apesar de todos os dados estarem disponíveis on-line, a visualização das informações é limitada caso-a-caso, o que torna inviável a coleta de dados manual. Portanto, enquanto não houver nenhuma divulgação sistemática de informações originais é impossível analisar a corrupção no Brasil utilizando esse conjunto de dados valiosos.

Este artigo aborda este problema através da introdução de um novo instrumento de coleta de dados automatizada. O TOGARY extrai as informações sobre as condenações de improbidade administrativa a partir do site do CNJ e exporta-as em formato de planilha. Uma vez que todos os dados são abertos e gratuitos, qualquer estudioso pode avançar a investigação científica sobre a corrupção no Brasil.

A partir de uma análise preliminar com os dados coletados, sabemos que o judiciário brasileiro leva mais de cinco anos, em média, para julgar um caso de corrupção. Além disso, identificamos que Bahia (8,86), Amazonas (8,77), Espírito Santo (7,99) e Alagoas (7,86) têm os sistemas judiciais mais lentos, enquanto Amapá (4,79), Mato Grosso do Sul (4,53), Tocantins (4,27), Minas Gerais (4,23) e Paraná (3,44) são os mais rápidos.

Além disso, os resultados empíricos também mostraram que não há diferença significativa entre os tribunais estaduais e federais em relação ao tempo de julgamento dos casos de corrupção. Em relação aos tribunais federais, constatou-se que a 5ª região é a mais rápida que as demais. No geral, as decisões tomadas por órgãos colegiados (3,8) são mais rápidas que as tomadas por trânsito julgado (5,74). Por fim, um outro achado significante é que, após a criação do CNJ, houve uma redução no tempo médio de julgamento dos casos de corrupção em todos os níveis de análise.

Com este trabalho, esperamos difundir a aplicação de procedimentos informatizados para coleta de dados em pesquisas empíricas de Ciência Política. Além disso, esperamos avançar o uso da ciência para aumentar a transparência pública e para combater a corrupção no Brasil.

REFERÊNCIAS BIBLIOGRÁFICAS

ADES, Alberto; DI TELLA, Rafael. (1999), Rents, competition, and corruption. American economic review, p. 982-993.

BRAUN, Miguel; DI TELLA, Rafael. (2004), Inflation, inflation variability, and corruption. *Economics & Politics*, v. 16, n. 1, p. 77-100.

BRUNETTI, Aymo; WEDER, Beatrice. (2003), A free press is bad news for corruption. *Journal of Public economics*, v. 87, n. 7, p. 1801-1824.

D'ORAZIO, Vito et al. (2014) Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis*, v. 22, n. 2, p. 224-242.

DOLLAR, David; FISMAN, Raymond; GATTI, Roberta. (2001) Are women really the "fairer" sex? Corruption and women in government. *Journal of Economic Behavior & Organization*, v. 46, n. 4, p. 423-429.

EASTERLY, William; LEVINE, Ross. (1997), Africa's growth tragedy: policies and ethnic divisions. *The Quarterly Journal of Economics*, p. 1203-1250.

FERNÁNDEZ-VILLAMOR, José Ignácio et al (2011). A semantic scraping model for web resources-Applying linked data to web page screen scraping.

FISMAN, Raymond; GATTI, Roberta. (2002), Decentralization and corruption: evidence across countries. *Journal of Public Economics*, v. 83, n. 3, p. 325-345.

FRIEDMAN, Eric et al. (2000), Dodging the grabbing hand: the determinants of unofficial activity in 69 countries. *Journal of public economics*, v. 76, n. 3, p. 459-493.

GEHLBACH, Scott. (2009), What Can Firm and Household Surveys Tell Us about Expert Assessments of Corruption?. *Paper presented at the APSA - American Political Science Association*.

GOLDEN, Miriam A; CHANG, Eric CC. (2007) Electoral systems, district magnitude and corruption. *British Journal of Political Science*, v. 37, n. 01, p. 115-137.

GRIMMER, Justin; KING, Gary. (2011), General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, v. 108, n. 7, p. 2643-2650.

GRIMMER, Justin; STEWART, Brandon M. (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, p. mps028.

GYLFASON, Thorvaldur. (2001), Natural resources, education, and economic development. *European Economic Review*, v. 45, n. 4, p. 847-859.

HARTMANN, B., Wu, L., COLLINS, K., & KLEMMER, S. R. (2007, October). Programming by a sample: rapidly creating web applications with d. mix. In *Proceedings of the 20th annual ACM symposium on User interface software and technology* (pp. 241-250). ACM.

HOPKINS, Daniel J.; KING, Gary. (2010), A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, v. 54, n. 1, p. 229-247.

JONG-SUNG, You; KHAGRAM, Sanjeev. (2005) A comparative study of inequality and corruption. *American Sociological Review*, v. 70, n. 1, p. 136-157.

KING, Gary; LOWE, Will. (2003), 10 million international dyadic events. *Dataverse Network*, Version, v. 3.

KUNICOVÁ, Jana; ROSE-ACKERMAN, Susan. (2005), Electoral rules and constitutional structures as constraints on corruption. *British Journal of Political Science*, v. 35, n. 04, p. 573-606.

LA PORTA, Rafael et al. (2000), Investor protection and corporate governance. *Journal of financial economics*, v. 58, n. 1, p. 3-27.

LAPALOMBARA, Joseph. (1994), Structural and institutional aspects of corruption. *Social research*, p. 325-350.

LEITE, Carlos A.; WEIDMANN, Jens. (1999), Does mother nature corrupt? Natural resources, corruption, and economic growth. Natural Resources, Corruption, and Economic Growth (June 1999). *IMF Working Paper*, n. 99/85.

MAURO, Paolo. (1995), Corruption and growth. *The quarterly journal of economics*, p. 681-712.

MONTINOLA, Gabriella R.; JACKMAN, Robert W. (2002), Sources of corruption: a cross-country study. *British Journal of Political Science*, v. 32, n. 01, p. 147-170.

MUNZERT, Simon et al. (2015), *Scraping the Web. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, p. 219-294.

PALDAM, Martin. (2001), Corruption and religion adding to the economic model. *Kyklos*, v. 54, n. 2-3, p. 383-413.

PAN, Alberto et al (2002). Semi-automatic wrapper generation for commercial web sources. In *Engineering Information Systems in the Internet Context* (pp. 265-283). Springer US.

PANIZZA, Ugo et al. (2001), Electoral rules, political systems, and institutional quality. *Economics and Politics*, v. 13, n. 3, p. 311-342.

PENMAN, Richard; BALDWIN, Timothy; MARTINEZ, David. (2009). *Web scraping made simple with sitescrape*.

POGGI, Nicólas et al (2007). Automatic detection and banning of content stealing bots for e-commerce. *In NIPS 2007 workshop on machine learning in adversarial environments for computer security*.

SANDHOLTZ, Wayne; KOETZLE, William. (2000), Accounting for corruption: Economic structure, democracy, and trade. *International studies quarterly*, v. 44, n. 1, p. 31-50, 2000.

SHELEIFER, Andrei. (1996), Origins of Bad Policies: Control, Corruption and Confusion. *Rivista di Politica Economica*.

SWAMY, Anand et al. (2001), Gender and corruption. *Journal of development economics*, v. 64, n. 1, p. 25-55.

TREISMAN, Daniel. (2000), The causes of corruption: a cross-national study. *Journal of public economics*, v. 76, n. 3, p. 399-457.

TREISMAN, Daniel. (2007), What have we learned about the causes of corruption from ten years of cross-national empirical research?. *Annu. Rev. Polit. Sci.*, v. 10, p. 211-244.

VARGIU, Eloisa; URRU, Mirko. (2012). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artificial Intelligence Research*, 2(1), p 44.