DOI: 10.5433/1981-8920.2025v30n4p156

A CONFIABILIDADE DE PESQUISAS JURIMÉTRICAS: DESAFIOS NA COLETA DE DADOS E ANÁLISES EMPÍRICAS DE JURISPRUDÊNCIA

THE RELIABILITY OF JURIMETRICS RESEARCH: CHALLENGES IN DATA COLLECTION AND EMPIRICAL CASE LAW ANALYSES

Alessandra Scherma Schurig[a]
Ana Claudia Batista[b]
Paulo Henrique Ferreira da Silva[c]
Daniel Oitaven Pearce[d]

RESUMO
Objetivo: evidenciar a necessidade de metodologia específica para validar a confiabilidade de pesquisas quantitativas na área jurídica, demonstrando como problemas da fase de coleta de dados prejudicam a validade e confiabilidade de uma pesquisa quantitativa. **Metodologia:** Análise exploratória com abordagem qualitativa e quantitativa**.** O artigo examinou trabalhos teóricos que estipulam como deve ser feita uma pesquisa de jurimetria e publicações normativas do Conselho Nacional de Justiça e do Supremo Tribunal Federal que regulam o tema, para compor diretrizes de importância, além de examinar sistemas externos de jurisprudência de oito tribunais estaduais. **Resultados:** são expostos estatisticamente problemas que dificultam sobremaneira ou até mesmo inviabilizam a pesquisa empírica quantitativa de jurisprudência calculando-se a qualidade da indexação utilizada nos mecanismos de busca de jurisprudência dos Tribunais Estaduais de São Paulo, Rio Grande do Sul, Paraná, Pernambuco, Minas Gerais, Distrito Federal, Rio de Janeiro e Amazonas. **Conclusões:** o texto debate teoricamente os problemas para a formação de uma cultura de precedentes que abarque pesquisas estatísticas e ainda demonstrou de modo quantitativo os problemas decorrentes da má qualidade de indexação e falta de uniformização de oito bancos de pesquisa de jurisprudência externos brasileiros, propondo possibilidades para a superação de dificuldades.

[a]Doutora em Direito pela Universidade Federal da Bahia (UFBA). Salvador, Brasil. E-mail: ale.scherma.schurig@gmail.com

[b]Doutoranda em Estatística e Ciências de Dados pela Universidade Federal da Bahia (UFBA). Salvador, Bahia. E-mail: anacsbatista87@gmail.com

[c]Doutor em Estatística pela Universidade Federal de São Carlos (UFSCAR). Docente na Universidade Federal da Bahia (UFBA). Salvador, Bahia. E-mail: paulohenri@ufba.br

[d]Doutor em Direito Público e em Ciências Sociais pela Universidade Federal da Bahia (UFBA). Docente na Universidade Federal da Bahia (UFBA). Salvador, Bahia. E-mail: danieloitaven@hotmail.com

Descritores: Análise de dados. Coleta de dados. Bancos de jurisprudência. Jurimetria. Uniformização.

1 INTRODUÇÃO

Uma pesquisa empírica se debruça acerca de fenômenos do mundo para tentar entendê-los através de observação e experimentação e sua principal diferença para os demais tipos de pesquisa é a sua caracterização através de uma pergunta de pesquisa com implicações observáveis a partir de um fenômeno concreto, possibilitando, desse modo, o estabelecimento de relações inferenciais.

Relações inferenciais têm por base o conceito de inferência, afeito à lógica na filosofia. Trabalhando junto à dedução e indução, a inferência possibilita a utilização de um conhecimento tácito, inferido do conhecimento explícito, já conhecido, ou seja, há um conhecimento declarativo, explícito, que dá origem a um conhecimento implícito, procedural, que pode ser explorado de diferentes maneiras.

Pesquisas empíricas, ao observarem fenômenos do mundo, descrevê-los e realizarem inferências a partir daquelas observações, trabalham com *dados,* compreendidos aqui como um outro termo para designar “fatos sobre o mundo”, que tomam formas variadas: podem ser históricos, contemporâneos, baseados em legislação ou jurisprudência, podem ser o resultado de entrevistas ou de pesquisas auxiliares arquivísticas.

Dados também podem ser precisos ou vagos, podem ser relativamente certos ou muito incertos, podem ser diretamente observados ou conseguidos indiretamente; podem ser antropológicos, sociológicos, econômicos, jurídicos, políticos, biológicos, físicos ou naturais. Desde que os fatos estejam de alguma maneira relacionados ao mundo, eles são dados, e, contanto que a pesquisa envolva dados que são observados ou desejados, ela pode ser classificada como empírica. (Epstein; King, 2013)

Existe uma diversidade de técnicas em pesquisas empíricas e, por conseguinte, uma diversidade de técnicas utilizáveis em pesquisas empíricas direcionadas ao sistema jurídico. Nesse sentido, uma das mais interessantes é

a pesquisa empírica do tipo *quantitativa*, que trabalha com a *técnica de análise de dados*, um processo que combina tecnologia e metodologias estatísticas para a composição de uma pesquisa que pode resultar em uma análise descritiva, ou até mesmo em uma análise preditiva, projetiva ou prescritiva.

Uma pesquisa empírica quantitativa trabalha com *medição*, o que envolve a comparação de um aspecto da realidade com um padrão, como os existentes para quantidades, capacidades e categorias (Epstein; King, 2013). Será a criação de medidas que permitirá que os dados de um fenômeno sejam analisados de modo mais compreensível e essas medidas devem ser garantidas por duas dimensões críticas: *confiabilidade* e *validade,* conforme explica Epstein e King (2013).

Confiabilidade é a extensão à qual se pode replicar uma medida, possibilitando a reprodução do mesmo valor, no mesmo padrão, para o mesmo tópico, a um mesmo tempo e pode ser interpretada como o inverso da variabilidade. Por sua vez, validade relaciona-se com o quanto uma medida confiável reflete o conceito fundamental sendo medido. Confiabilidade e validade são parâmetros que possibilitam a replicação de uma pesquisa por outros pesquisadores, garantindo a correção das informações.

Através de medição e da garantia de confiabilidade e validade, a análise de dados pode criar meios muito eficientes para relatar e descrever um fenômeno que dependa de uma quantidade de dados muito grande e complexa, que à primeira vista não forneceria informações compreensíveis. Para uma conjuntura de grandes quantidades de dados, uma das técnicas mais promissoras é a mineração de dados, uma intersecção de estatística, inteligência artificial (IA), aprendizado de máquinas e gerenciamento de dados que constrói modelos de base matemática para a exibição de padrões exploratórios, prescritivos ou preditivos que podem ser encontrados nas bases analisadas.

O sistema jurídico pode se beneficiar com esse modelo de pesquisa para observar e analisar quantitativamente direções estabelecidas por diferentes objetos jurídicos e, por essa via, auxiliar determinada compreensão acerca de fenômenos que impactem o direito. É válido, contudo, observar que os dados e os resultados de suas análises nada dizem *per si*, pois exigem interpretação e

uma base teórica sólida que forneça fundamentos para essa interpretação. Essa é uma fase fundamental da pesquisa empírica com dados e uma fase que depende de experiência e conhecimento sólido sobre a área de produção dos dados que devem ser aliadas ao conhecimento sobre estatística e suas possibilidades de medição.

Tais imperativos de conhecimento evidenciam que uma pesquisa empírica quantitativa jurídica exige tipos de conhecimento distintos, que vão desde conhecimento jurídico até conhecimento estatístico, em especial quando se realiza *análise de dados jurisprudenciais.* Nesse modelo de pesquisa, a pretensão é realizar análise quantitativa de um conjunto de dados de jurisprudência para encontrar respostas para um questionamento específico ou mesmo para observar o que aqueles dados podem informar, sem que nenhuma pergunta específica seja feita de antemão. Estudos desse tipo podem ser denominados como *estudos de jurimetria*, denominando o emprego de técnicas de estatística ao universo jurídico. Decisões judiciais se tornam o objeto principal da pesquisa, que pode ser direcionada para variadas vertentes: órgãos prolatores, temas específicos, períodos temporais e assim por diante.

Em regra, uma pesquisa começa com uma pergunta. Ao analisar os dados para uma possível resposta, é possível encontrar outras informações que desafiam o interesse inicial do pesquisador. Por exemplo, ao se realizar uma pesquisa jurimétrica, pode-se começar indagando sobre a quantidade de condenações por improbidade administrativa em determinado tribunal. No decorrer da análise dos dados, pode se perceber uma grande quantidade de decisões terminativas, o que conduziria à análise da peça inicial, sendo possível apontar incorreções na sua confecção, por exemplo e assim descrever problemas e propor soluções para uma petição inicial que tenha melhores chances de sucesso quando analisada pelo órgão judicial. Como visto através do exemplo, esse tipo de movimento ampliativo em uma pesquisa não deve ser rejeitado, até porque traz consigo potenciais desdobramentos interessantes para a pesquisa.

Não obstante comportar essa flexibilidade em seus movimentos de pesquisa, a análise de dados jurisprudenciais deve cumprir uma série de

requisitos para garantir a legitimidade de sua metodologia em diferentes etapas que abrangem as fases necessárias para o estudo metodologicamente correto e válido do fenômeno, garantindo a confiabilidade e a validade das pesquisas realizadas.

Discorrendo sobre as etapas fundamentais da pesquisa jurimétrica, cite-se que a primeira fase é a *coleta e armazenamento de dados*. Aqui serão buscadas e examinadas fontes que contenham os dados de interesse para a pesquisa, que devem ser coletados e armazenados para que possa passar à segunda fase, o *processamento,* quando os dados coletados são preparados, organizados e resolvidos problemas com a qualidade, limpeza e agregação, sendo realizada a consolidação dos dados. Com dos dados consolidados, passa-se para a última fase: a *análise propriamente dita*, quando são utilizadas técnicas estatísticas e teorias jurídicas para interpretar os dados e construir respostas para as hipóteses e questionamentos da pesquisa, inclusive sobre desdobramentos trazidos pela própria análise de dados que sejam mais amplos ou distintos dos questionamentos iniciais.

Cada etapa apresenta seus próprios desafios: por exemplo, a fase de processamento lida com a constante possibilidade da presença de *ruídos* nos bancos de dados que podem prejudicar toda a análise a ser realizada por conta de contaminação. É por isso que os dados precisam ser tratados, mediante a aplicação de diferentes métodos, permitindo a identificação e correção de dados para a padronização do que foi colhido para possibilitar a análise. Outra fase repleta de desafios é a fase de interpretação, que depende diretamente da presença de conhecimento de qualidade sobre a área analisada, pois é a presença de uma sólida base de conhecimento que possibilitará a formação de uma boa pergunta de pesquisa e a análise correta dos desdobramentos dos dados analisados. Além disso, é necessário bom conhecimento estatístico, pois é essa ciência que permitirá entender o que pode ou não ser encontrado e respondido através de medições.

Entretanto, todas essas etapas dependem essencialmente da existência e boa condução da primeira fase: a *fase de coleta de dados,* fase que apresenta problemas específicos que serão relatados a seguir.

2 DISCUTINDO PROBLEMAS DA FASE DE COLETA DE DADOS EM PESQUISAS QUANTITATIVAS DE JURISPRUDÊNCIA

A fase de coleta e armazenamento de dados para uma pesquisa é a fase em que são levantadas as fontes dos dados e verificada a possibilidade de extração daqueles dados. E, nesse ponto, surge uma pergunta fundamental: *quais dados devem ser coletados?* Para responder essa pergunta, alguns passos são essenciais: (1) identificar a população de interesse; (2) coletar o máximo de dados possível com base em alguma metodologia de amostragem; (3) registrar o processo pelo qual os dados foram observados; e (4) coletar dados de uma maneira que reduza o viés de seleção (Epstein; King, 2013).

Essas respostas irão variar de acordo se os dados coletados são *primários*, ou seja, obtidos diretamente pelo pesquisador que planejou a coleta e formulou as hipóteses do estudo. Isso pode ocorrer através de observações, entrevistas e outros métodos. Por outro lado, pode ser o caso do trabalho com dados *secundários*, coletados por outras pessoas que não são os pesquisadores daquela pesquisa. Essas são situações que influenciam na decisão sobre quais dados devem ser coletados.

Na fase de coleta é importante prestar atenção na *quantidade de dados* que devem ser coletados para serem analisados, sendo esse ponto um indicativo de uma boa pesquisa, apesar da quantidade não representar em si qualidade, mas é necessária a busca por uma quantidade mínima, representativa da população. Em quase toda utilização empírica concebível, uma maior quantidade de dados não prejudica os objetivos do pesquisador, tornando sempre sua amostra mais próxima da população estudada (Epstein; King, 2013).

Por vezes existirá a impossibilidade de coleta de todos os membros de uma população. Assim, a quantidade de dados a serem coletados deve ser em número suficiente para a pesquisa. Surge então a importância da *amostra:*

> […] se as circunstâncias impedem que os pesquisadores coletem dados de todos os membros da população, mas eles possuem recursos para coletar um grande número de observações, eles devem estabelecer uma amostra de probabilidade aleatória – uma amostra em que cada elemento na população total tem uma conhecida (e preferencialmente a mesma) probabilidade de ser selecionado. (Epstein; King, 2013,

p. 139).

A técnica estatística de amostragem garantirá a extrapolação das informações amostrais para a população de estudo com certo nível de confiança. Em regra, uma análise de dados apresenta *elementos*, uma *população* e uma *amostra.* Os elementos são as partes de um conjunto, como por exemplo, lápis de cor que façam parte de uma caixa de lápis. A população é o conjunto de elementos que será estudado, por exemplo, a caixa de lápis de cor. E a amostra é um subconjunto representativo da população, ainda trabalhando com o exemplo, podem ser todos os lápis de cores primárias daquela caixa.

A técnica de amostragem pretende possibilitar o estudo de somente uma seleção daquela população e permitir que os resultados observados sejam válidos para o conjunto, através de uma extrapolação que pode ser feita mediante o uso de técnicas de inferência estatística como intervalos de confiança ou testes de hipóteses, por exemplo.

É essencial também prestar atenção na *qualidade dos dados*, que devem respeitar as regras de uma coleta ética e rigorosa. Para tanto, observar as *fontes dos dados* é primordial, verificando se são levantamentos amostrais ou não, se cobrem corretamente o universo de pesquisa, se são bases de dados públicas ou privadas, quais as formas de obtenção daqueles dados, se há obediência à legislação de regência e assim por diante.

Existindo dificuldades na extração direta de dados, o próximo passo poderá ser o pedido de *compartilhamento dos dados*, o que pode ser feito através de requisições enviadas aos órgãos jurisdicionais responsáveis, que podem ou não responder favoravelmente de acordo com seus critérios particulares. O outro passo possível é realizar a pesquisa nos *sistemas de busca online de jurisprudência.*

Quando um pesquisador externo aos tribunais faz uma pesquisa quantitativa de análise de jurisprudência, ele trabalha com dados secundários, porque os dados já foram coletados, indexados e organizados pelos respectivos tribunais em sistemas de organização, um dos quais é o sistema de busca online de jurisprudência, uma base de dados que reúne os julgamentos de determinado tribunal, agrupados por diversos critérios. Esses bancos de jurisprudência

podem ser internos, de acesso restrito aos funcionários daquele tribunal, compostos por dados que não são de acesso público. Existe ainda uma espécie de “espelho” dessa base interna, o sistema de jurisprudência externo, o citado sistema de busca online, uma interface de busca que reúne os julgamentos de determinado tribunal para pesquisa por usuários externos, possibilitando pesquisas textuais nos documentos disponibilizados e organizados.

É importante entender que existem diferenças nas duas bases de dados, externas e internas, citadas, sendo necessário destacar que as bases de busca de jurisprudência externa *não refletem fidedignamente todos os julgados de um tribunal em certo período temporal*, disponibilizando apenas parte dos processos julgados, pois existe a seleção das secretarias dos tribunais, baseada em critérios que nem sempre são claros:

> Nesse sentido, as ferramentas de pesquisa de jurisprudência disponíveis nas páginas eletrônicas dos tribunais pesquisados não pareciam ser suficientes para que os usuários da Internet (principalmente aqueles que não possuem o domínio da pesquisa avançada) pudessem ter acesso pleno às decisões. Além das dificuldades no uso das ferramentas de busca, os autores também perceberam que os bancos de dados online de jurisprudência não continham todos os julgados. Portanto, os julgados disponíveis ao público para acesso por meio da Internet consistiam em apenas uma amostra do total de casos decididos pelos tribunais – uma amostra cujos critérios de escolha não estavam explícitos e que, portanto, não garantiam ao usuário que o resultado de sua pesquisa online representasse fidedignamente a posição dos tribunais sobre o tema pesquisado (Veçoso, *et al.*, 2014, p. 110).

Ou seja, o que está nas bases de dados de jurisprudência externas já são amostras dos julgados, agrupados de acordo com critérios decididos por cada tribunal. É importante deixar isso claro: as bases externas não demonstram a realidade da quantidade de julgados em determinado período temporal, como por exemplo, no STF, onde a base de dados não é exaustiva porque contém apenas uma amostra das decisões colegiadas e mais, os critérios de seleção para decisões monocráticas são variáveis.

E o ponto aqui é o seguinte: a dificuldade em precisar quais são os critérios de seleção para disponibilizar aquele julgado para consulta externa cria um problema para a confiabilidade dos resultados. Assim, seria importante o acesso à completude dos dados jurisprudenciais em determinado segmento

pesquisado, sendo possível, de alguma forma, a disponibilização de todas as decisões judiciais para pesquisa, conforme é explicado pelos pesquisadores no trecho abaixo:

A constituição do banco de dados integral se mostra relevante na medida em que possibilita extrair informações seguras e efetivamente representativas da produção jurisprudencial. Isso impede qualquer tipo de viés na seleção das decisões que serão acessíveis para pesquisa online e possibilita ao pesquisador selecionar sua própria amostra de pesquisa, definindo a informação que considera relevante para sua pesquisa, de acordo com os parâmetros e com a confiabilidade que ele mesmo quer estabelecer para sua pesquisa. A completude do banco de dados é de especial importância para pesquisas quantitativas, em que a definição de uma amostra representativa, confiável e não enviesada é essencial. (Veçoso, *et al.,* 2014)

Os tribunais alegam que disponibilizar todos os julgados seria contraproducente, prejudicando os sistemas e a agilidade da busca nos mesmos, pois haveria a repetição de milhares de documentos no mesmo sentido. Esse é um bom argumento, porém, o que se pede é que, para fins de análises mais aprofundadas, se torne possível a disponibilização de bases de dados completas, sem tratamento prévio, pois, muitas vezes, o objeto da pesquisa incide justamente em identificar o padrão contido em julgados semelhantes. Portanto, esse tratamento prévio já estaria enviesando os achados da pesquisa.

No Brasil, os sistemas de busca são formados e organizados de diferentes maneiras e de acordo com regras internas de cada tribunal, padecendo de uniformização. Esse é um problema que o Conselho Nacional de Justiça (CNJ) vem tentando solucionar nos últimos anos através de diversas propostas. Nesse sentido, é possível citar a criação do Banco Nacional de Precedentes (BNP), instituído pela Resolução do CNJ de número 444 de 25 de fevereiro de 2022, desenvolvida para permitir a consulta e divulgação de precedentes judiciais em uma base unificada. Esta base está em funcionamento, mas ainda em desenvolvimento para agregar mais funções e mais utilidades. Há ainda o importante Pangea/BNP, uma ferramenta pública destinada a pesquisadores e

desenvolvida pelo Tribunal Regional do Trabalho da 4ª Região para facilitação do uso dos diversos precedentes qualificados utilizados na jurisdição. Apesar dessas iniciativas, não existe ainda uniformização nos sistemas de busca de jurisprudência e longe de ser tema dispensável, esse é um problema que merece atenção.

Nos últimos anos o Brasil vem desenvolvendo uma cultura de respeito aos precedentes e a popularização de pesquisas e análises desses precedentes é um passo importante para a consolidação dessa cultura que ainda é nova em uma país de tradição de *civil law.* E a ausência de uniformização nas bases de dados jurisprudenciais causa problemas para a realização de estudos de qualidade, muitas vezes empurrando o pesquisador para serviços pagos da iniciativa privada, que através de táticas de manipulação, conseguem acessar e agrupar os dados que deveriam ser disponibilizados publicamente sem maiores problemas[1].

A realização de uma pesquisa já depende intrinsecamente da qualidade do pesquisador e de seus esforços e a falta de uniformização nas bases de jurisprudência força os pesquisadores a conhecer e examinar cada sistema e suas peculiaridades para posteriormente descrever essas peculiaridades na sua pesquisa e trabalhar com elas para a análise de dados, sendo que a cada fonte, diferentes desafios serão apresentados.[2]

1 Uma boa resposta para a necessidade de uniformização dos sistemas brasileiros foi criada em 2020 com a DataJud (Base Nacional de Dados do Poder Judiciário), iniciada através da Resolução Nº 331 de 20/08/2020 e instituída como fonte primária de dados do Sistema de Estatística do Poder Judiciário (SIESPJ). A DataJud é composta por dados e metadados processuais relativos a todos os processos físicos ou eletrônicos, públicos ou sigilosos, de qualquer das classes previstas nas Tabelas Processuais Unificadas (TPU), que por sua vez objetivam a padronização e uniformização taxonômica e terminológica de classes, assuntos, movimentação e documentos processuais no âmbito da Justiça Estadual, Federal, do Trabalho, Eleitoral, Militar da União, Militar dos Estados, do Superior Tribunal de Justiça e do Tribunal Superior do Trabalho, a serem empregadas em sistemas processuais. O CNJ determinou que os tribunais organizassem e desenvolvessem seus sistemas internos, inclusive para organizar os processos que já estivessem baixados, visando, portanto, uma organização geral dos arquivos de cada tribunal. A ideia é que peças, documentos e decisões sejam cadastrados de modo uniforme e, de fato, através dessas medidas, tenta-se viabilizar, ao menos desde 2020, alguma uniformidade.

2 Por exemplo, no Supremo Tribunal Federal (STF), a fonte primária para construção da base de dados foi o Ementário de Jurisprudência do STF, coleção iniciada em julho de 1950 e composta pelos textos integrais dos acórdãos publicados a partir de então. Depois incorporou-se documentos da chamada COLAC (Coletânea de Acórdãos), organizada a partir de 1982 com foco nos acórdãos publicados entre 1936 e 1949, isto é, nos acórdãos anteriores à criação do

Segundo a determinação do CNJ, os sistemas internos de jurisprudência dos tribunais deveriam ser alimentados e organizados através das TPU-Tabelas Processuais Unificadas. Espera-se, ainda segundo as determinações do CNJ, que os tribunais adaptem seus sistemas internos e implementem as TPU. A partir da data de implementação das TPU, aguarda-se que os novos processos já sejam cadastrados de acordo com as tabelas unificadas de classes e assuntos processuais. Isso deveria facilitar as pesquisas internas e externas e produzir uniformização, sendo um passo indispensável para, por exemplo, pesquisadores voltados a estudar a jurisprudência de certo tribunal e que pretendam o compartilhamento de dados, bastariam que eles indicassem os dados de interesse de acordo com as TPU. Mas como ainda não existe implementação na maioria dos tribunais, não há uniformização próxima.

Desse modo, enfrentando problemas no compartilhamento, falta de uniformização nas bases de jurisprudência e problemas em agregar diferentes tipos de conhecimento, a pesquisa quantitativa na área jurídica enfrenta dificuldades imprevistas, sendo que o acesso aos dados jurisprudenciais para pesquisas através de estudos estatísticos não deveria enfrentar dificuldades. Porém, o acesso a esses dados para análise em pesquisa é uma situação que enfrenta muitos percalços e a pesquisa quantitativa de jurisprudência, já rara em nosso país, nasce enfrentando dificuldades sem fim para sua realização, situação que em mais de uma década, não tem encontrado solução:

> O acesso à informação é um aspecto da accountability das instituições públicas que tem reflexos na disponibilização das bases eletrônicas de julgados, na medida em que estas permitem a análise das decisões judiciais. Contudo, essa ainda é uma política em construção nos tribunais do país (Veçoso *et al.*, 2014, p. 107).

Em sua gênese, dados processuais são dados públicos. A Constituição Federal, em seu artigo 5º, inciso LX, determina que a lei só poderá restringir a publicidade dos atos processuais quando a defesa da intimidade ou o interesse social o exigirem, garantindo, portanto, uma restrição ao direito quando necessário o sigilo; mas a regra é a publicidade dos atos processuais. Também

Ementário. Hoje, o banco do STF abrange julgados desde julho de 1950 até o período mais atual.

o Código de Processo Civil, em seu artigo 189, dispõe que os atos processuais são públicos, restringindo essa publicidade a hipóteses de necessidade de resguardo.

É plenamente possível aos tribunais resguardar os dados protegidos por sigilo processual, ou seja, aqueles dados que seguem o quanto disposto sobre a necessidade de sigilo pela Constituição Federal e demais legislações, como por exemplo o novel microssistema de proteção às mulheres, dos dados públicos, que não devem possuir restrição de privacidade, já que se trata de informação aberta que pode ser analisada e compartilhada. Portanto, essencialmente, não existem fundamentos jurídicos que proíbam o acesso a dados de jurisprudência. O que existe atualmente é um quadro de ineficiência no cumprimento do anteparo normativo, que vai desde a Constituição até as resoluções do CNJ.

3 DESENVOLVENDO UMA METODOLOGIA POSSÍVEL PARA A COLETA DE DADOS EM PESQUISAS QUANTITATIVAS DE JURISPRUDÊNCIA QUE DEPENDAM DE SISTEMAS DE BUSCA ONLINE

A primeira e maior dificuldade em relação a pesquisas quantitativas que necessitem de dados de jurisprudência e façam consultas em sistemas de busca online é a possibilidade de acesso aos dados de jurisprudência de forma estruturada. A não estruturação dos dados dificulta sobremaneira uma pesquisa quantitativa e essa dificuldade de exportação de resultados para análise estatística é um tema que precisa ser discutido pela comunidade de pesquisa científica.

Para esclarecer, *dados estruturados* são aqueles organizados em formas tabulares, em linhas e colunas, com colunas para cada variável e restrições de tipo e formatação de valores de cada coluna. Em consulta realizada em 2024 pelos autores nos sítios de pesquisa jurisprudencial dos tribunais do país, verificou-se que além do STF, somente o Tribunal de Justiça de Pernambuco permite o download completo em arquivo CSV (comma-separated-values) com ementa e decisão completa e o Tribunal de Sergipe permite apenas um download de CSV por página. Até mesmo o Banco Nacional de Precedentes,

Pangea/BNP, apesar de unir diversos tribunais e diversos precedentes qualificados, não permite a exportação dos dados de forma estruturada.

Ora, atualmente há uma mudança importante que justifica esse acesso a dados estruturados, não apenas pela importância da construção da ideia da cultura de precedentes, mas também em razão do advento da jurimetria e do uso de técnicas estatísticas para análise de dados jurídicos, da chegada do Big Data e das técnicas de Processamento de Linguagem Natural, esse conteúdo passou a ser de interesse de um novo nicho de pesquisadores: estatísticos, cientistas de dados e juristas que possuam formação empírica em pesquisa.

Assim, se antes dados oriundos da jurisprudência eram inicialmente de interesse, particularmente, de advogados, promotores e demais pessoas envolvidas na área do direito e por isso o mecanismo de busca e extração de informações individuais, tal como proposto pelos tribunais, era suficiente, esse novo cenário desafia o acesso e disponibilidade de dados brutos, em larga escala e variedade. Tal acesso tornou-se objeto de grande valor, com o potencial de gerar conhecimento valioso para a área jurídica.

A ausência de compartilhamento de dados estruturados e a falta de uniformização nos sistemas de busca dos tribunais compõem uma etapa complexa da pesquisa quantitativa, situação que não precisaria existir. O ponto é que a falta de estruturação dos dados atrapalha a coleta em virtude da quantidade de julgados que pode ser encontrada ao se pesquisar um termo como, por exemplo, “dolo e improbidade e administrativa”, com filtro de período do julgamento entre 01 de janeiro de 2019 e 31 de dezembro de 2023, conforme revelado no Quadro 1:

Quadro 1 - Relação de julgados por tribunal e disponibilidade de dados estruturados

ÓRGÃO	ACÓRDÃOS	DOWNLOAD	OBSERVAÇÃO
STJ	726	Não permite	Coleta de informação individual
STF	366	Permite	Um arquivo estruturado completo em planilha eletrônica

TJSP	6.151	Não permite	Coleta de informação individual
TJMG	3.985	Não permite	Coleta de informação individual
TJPR	2.695	Não permite	Coleta de informação individual
TJRS	1.270	Não permite	Coleta de informação individual
TJRJ	560	Não permite	Coleta de informação individual
TJPE	731	Permite	Um arquivo estruturado completo em planilha eletrônica
TJAM	212	Não permite	Coleta de informação individual
TJDFT	3.933	Não permite	Coleta de informação individual
TJBA	-	Não permite	Coleta de informação individual

STJ – Superior Tribunal de Justiça; STF – Supremo Tribunal Federal; TJSP – Tribunal de Justiça de São Paulo; TJMG - Tribunal de Justiça de Minas Gerais; TJPR - Tribunal de Justiça do Paraná; TJRS - Tribunal de Justiça do Rio Grande do Sul; TJRJ - Tribunal de Justiça do Rio de Janeiro; TJPE - Tribunal de Justiça de Pernambuco; TJAM - Tribunal de Justiça do Amazonas; TJDFT - Tribunal de Justiça do Distrito Federal; TJBA - Tribunal de Justiça da Bahia

Fonte: elaboração própria (2024).

Sendo assim, a importância desse tipo de exportação de dados, de forma estruturada, para a formulação das pesquisas é fundamental. Em regra, o que se deseja é, ao menos, um arquivo em formato de planilha eletrônica, que é um arquivo de texto com campos de dados separados por vírgulas, o que facilita o planilhamento e análise. É um modelo de arquivo compatível com múltiplos aplicativos e de fácil utilização.

Não obstante, os dados podem ser estruturados de formas distintas do modelo de planilha eletrônica sem grandes problemas, mas há dificuldade na implementação dessa alternativa nos bancos de dados de jurisprudência do país disponibilizados para consulta externa, onde os dados estão sem estruturação ou sem método de exportação. Desse modo, pesquisadores empíricos reclamam regularmente que a obtenção de dados é tarefa complicada, sendo “[...] bastante comum a necessidade de construção de ferramentas computacionais complexas para tornar a tarefa possível, ferramentas estas que podem se tornar obsoletas

com simples atualizações dos sistemas". (Trecenti, 2016, p. 21).

O pesquisador Trecenti (2016) relata que os tribunais ou disponibilizam ferramentas que precisam do número identificador do processo ou que as ferramentas de consulta de jurisprudência são duvidosas. Por isso, ele se voltou a desenvolver estratégias para a coleta de dados estruturada e eficiente, desenvolvendo técnicas de raspagem de documentos com uso de softwares livres, como por exemplo o R, disponibilizando os códigos para acesso público por outros pesquisadores. Seu programa é uma espécie de robô que imita ações do ser humano, acessando páginas, preenchendo formulários e realizando as ações necessárias para receber documentos que são armazenados localmente para serem processados posteriormente.

O programa de extração de Trecenti resolve a situação de estruturação dos dados porque consegue armazenar os resultados em páginas HTML *(HyperText Markup Language)* ao passar por uma fase de raspagem dos documentos, transformando os dados em formato semiestruturado em dados estruturados e pode ser adaptado para diversos tribunais, sendo ainda possível extrair dados de diários oficiais[3]. Como é comum com pesquisadores estatísticos e cientistas de dados, essas ferramentas são disponibilizadas para uso público, o que ajuda na solução do problema, mesmo que este problema não tenha fundamento para existir pois, se o dado é público e está sendo exibido em uma ferramenta de busca externa, então existe um Sistema de Gerenciamento de Banco de Dados (SGBD) interno que o alimenta, pois, de todo SGBD é possível exportar dados estruturados nas mais diversas extensões.

Uma alternativa para a impossibilidade de acesso a dados estruturados poderia ser a hipótese de utilização dos diários de justiça eletrônicos dos tribunais. Mas se esbarra aqui novamente ma falta de uniformidade, pois cada tribunal disponibiliza o seu Diário de Justiça Eletrônico (DJE).

A hipótese de realizar a coleta de dados nos diários oficiais também é problemática, principalmente porque não há publicação em formato estruturado

[3] TRECENTI, Julio Adolfo Zucon. Diagramas de influência: uma aplicação em Jurimetria. 2016. Dissertação (Mestrado) – Universidade de São Paulo, São Paulo, 2016. Disponível em: https://teses.usp.br/teses/disponiveis/45/45133/tde-20230727-113325/. Acesso em: 04 jun. 2024. Fls 23

e sim em páginas PDF *(Portable Document Format)*, sendo possível fazer o download dos cadernos em alguns diários eletrônicos somente, o que novamente desafiará a criação de ferramentas para superação desse óbice ou o que é lamentável, a utilização de sistemas pagos que possibilitem o acesso aos arquivos de forma estruturada, lucrando com uma informação que deveria ser pública e mais, por conta de problemas de opacidade técnica ou de proteção, podem nem ao menos ser confiáveis.

Quando se examinam as peculiaridades sobre definições conceituais, há que se atentar para a possibilidade de buscar auxílio com o uso dos *Tesauros,* que são listas de palavras controladas, vocabulários organizados que lidam com as complexidades da linguagem natural, compondo instrumentos de busca e acesso à informação disponibilizados por cada tribunal e, muitas vezes, fazendo conexão direta com os bancos de jurisprudência:

> Tesauros como instrumentos de organização do conhecimento, ou melhor, como linguagens documentárias utilizadas no processo de indexação, são listas estruturadas de termos e suas relações, onde cada um deve representar um único conceito ou ideia, de forma a orientar indexadores e usuários, levando-os de uma ideia ao termo que melhor a expresse. (Pinheiro; Ferrez, 2014, p. 9)

Os Tesauros podem ser usados em qualquer domínio de conhecimento, e são elaborados por especialistas que procuram acompanhar a atualização dos assuntos daquele domínio.[4] Na área jurídica existem diversos Tesauros, como o Tesauro do Tribunal Superior Eleitoral, o Tesauro da Justiça Federal, o Tesauro do Supremo Tribunal Federal e o Tesauro do Superior Tribunal de Justiça, dentre outros, acessíveis para consulta externa e que podem ser bastante úteis para a definição dos termos a serem pesquisados externamente.

Seguindo com a pesquisa ainda na primeira fase de coleta de dados nesse banco online de jurisprudência, um passo importante é prestar atenção

[4] A ideia de um Tesauro é melhorar o acesso à informação através da organização do conhecimento para modelagem de um domínio específico, sendo que sua utilização pode auxiliar tanto na recuperação de documentos ao sistematizar os conceitos e as relações entre esses conceitos, como também ajudar na própria indexação de documentos. Sua estrutura é desenhada de forma a demonstrar relacionamentos e ligações entre conceitos que são representados por determinado termo, sendo que esse termo deve ser sempre ligado a outro para compor relacionamentos. Esses relacionamentos podem ser hierárquicos, ontológicos, de equivalência, como sinônimos ou homônimos, ou relacionamentos de efeito, que ocorrem quando há causalidade ou descendência entre termos.

nos numerosos campos que a página de pesquisa de cada tribunal disponibiliza. Isso porque, em regra, esses campos refletem *modos de indexação*, então é necessário verificar se a pesquisa utilizará data da publicação do julgado ou data de julgamento, por exemplo. Ainda é necessário entender que tipo de origem terá aquele julgado: colegiado ou singular? O desejo é pesquisar pela ementa ou pelo inteiro teor? Todos esses são pontos que devem ser relatados na metodologia da pesquisa realizada que precisa descrever as definições conceituais e termos através do uso do Tesauro ou por outro método, registrar os diferentes usos de conectores ou operadores de pesquisa e atentar para a disponibilidade de cada sistema de busca.

Contudo, a utilização de Tesauros auxilia, mas não sana o problema da impossibilidade de acesso a dados estruturados. Nesse sentido, a gravidade do problema da ausência de possibilidade de exportação de dados estruturados pode ser visualizada com o exemplo a seguir. Realizando-se a pesquisa do termo “improbidade” nas páginas do Tesauro do Superior Tribunal de Justiça é alcançado o resultado mostrado na Figura 1.

Figura 1 - Tesauro do Superior Tribunal de Justiça, busca pelo termo “improbidade”

IMPROBIDADE ADMINISTRATIVA		
	TR	AGENTE POLÍTICO
	TR	ASSESSOR PARLAMENTAR
	TR	LEI DA FICHA LIMPA
	TR	NEPOTISMO
	TR	PRINCÍPIO DA MORALIDADE
	TR	PRINCÍPIO DA PROBIDADE ADMINISTRATIVA
	TR	PROBIDADE ADMINISTRATIVA
	TR	PROGRAMA NACIONAL DE ALIMENTAÇÃO ESCOLAR (PNAE)
	TR	SECRETÁRIO DE OBRAS E URBANISMO
	CAT	ADM/DAG, ADM/DAB, DC/DCTC
IMPROBIDADE (DIREITO DO TRABALHO)		
	NOTA	CONSTITUI JUSTA CAUSA PARA RESCISÃO DO CONTRATO DO TRABALHO PELO EMPREGADOR - ART. 482 DA CLT.
	TR	DESPEDIDA POR JUSTA CAUSA
	TR	EMPREGADO
	TR	EMPREGADOR
	TR	JUSTA CAUSA
	CAT	DC/DCTB

2 termos principais encontrados.

Fonte: Superior Tribunal de Justiça disponível em STJ - Jurisprudência do STJ (2024)

Por outra via, a busca do mesmo termo, "improbidade", através do Tesauro do STF, dará margem aos resultados presentes na Figura 2:

Figura 2 - Tesauro do STF, busca pelo termo “improbidade”

Fonte: Supremo Tribunal Federal (2024)

Note-se a primeira dificuldade: o termo pesquisado é o mesmo, mas a organização do Tesauro é diferente. A próxima dificuldade será a quantidade considerável de julgados que será apresentada como resultado, somada à impossibilidade de exportação dos dados estruturados. Ao usar o Tesauro do STJ e selecionar o termo “improbidade administrativa”, em junho de 2024 o pesquisador será direcionado para o banco de dados de jurisprudência deste tribunal sendo apresentado ao resultado de 6.526 acórdãos, 8 súmulas, 201 informativos e 65.960 decisões monocráticas [5]. Enquanto no STF, selecionando o termo em caixa alta “improbidade administrativa”, há o direcionamento para o banco de jurisprudência do tribunal contendo sobre o termo 854 acórdãos, 7.792 decisões monocráticas, 81 informativos, 21 temas de repercussão geral e 10 questões de ordem[6].

Diante da magnitude de resultados, o que o pesquisador deve fazer? Como selecionar dados de seu interesse? Ele deverá ler todos os acórdãos e verificar os argumentos desenvolvidos? Tudo isso é possível, mas não são técnicas de análise de dados. O que é recomendável em uma pesquisa quantitativa é utilizar amostras, pois pesquisas que utilizam amostragem lidam com um número substancial de dados para serem analisados e a amostra

5 De acordo com pesquisa realizada pelos autores em STJ - Jurisprudência do STJ, acesso em 10 de junho de 2024, às 10:50h.

6 De acordo com pesquisa realizada pelos autores em Pesquisa de jurisprudência - STF, acesso em 10 de junho de 2024, às 10:53h.

garante que o trabalho pode ser feito analisando uma quantidade menor, com a garantia que a qualidade do trabalho não se perderá, evitando que dados isolados sejam transformados em dados representativos.

Mas, imagine-se que o pesquisador tenha selecionado dados de acordo com o que ele considera que são argumentos bem desenvolvidos e publique uma pesquisa afirmando que aquele conjunto de dados representa o entendimento do STF sobre improbidade. Isso está tecnicamente incorreto, porque este pesquisador está selecionando subjetivamente julgados de interesse particular e essa escolha não representa o entendimento de um tribunal.

Por outro lado, metodologicamente, uma amostra calculada segundo os critérios matemáticos corretos pode sim representar o entendimento de um tribunal porque o uso de amostras de probabilidade aleatória garante a possibilidade de extrapolação das análises para a população, com certo nível de confiança, bem como a ausência de associações entre regras de seleção e variáveis. Por isso, recomenda-se que sejam usadas em estudos com população volumosa, como, por exemplo, o caso apresentado de análise de jurisprudência através da busca por termos. Nada impede que após a pesquisa quantitativa, o pesquisador opte por selecionar julgados, em uma etapa qualitativa da pesquisa e, então, aprofunde-se no exame e debate sobre pontos específicos.

Contudo, aqui também o pesquisador irá se deparar com o problema da uniformização, visto que coletar uma amostra é um processo de metodologia rigorosa que irá enfrentar a “[...] total falta de uniformidade nos mecanismos de busca utilizados pelos diferentes tribunais do país.” (Veçoso *et al.*, 2014, p. 109), pois cada tribunal pode empregar e fornecer os métodos que considerar necessários e suficientes, sem uniformidade.

Seria possível aos tribunais disponibilizar um link para API (Application Programming Interface – Interface de Programação de Aplicações) de acesso aos dados, ou então facilitar o contato com o suporte técnico de TI (Tecnologia da Informação) responsável pelo SGBD de jurisprudência. Essas seriam soluções simples que atenderiam os pesquisadores da área. Então, esses documentos são agrupados, e é importante destacar que em alguns tribunais

são agrupados com a utilização de ferramentas de IA[7] [8]

Assim, existem problemas que que impactam a qualidade nos bancos de jurisprudência e são problemas que não são solúveis apenas com a disponibilização de painéis estatísticos pelo CNJ, apesar desta ser uma iniciativa louvável e de grande valia. Mas ali também não há possibilidade de download dos dados estruturados e não há acesso à integralidade das decisões, apenas aos números brutos envolvendo informações de interesse, como quantidade de processos julgados, incidência do assunto por tribunal. Mas não há acesso ao julgado e é fundamental para uma certa vertente de pesquisas quantitativas o acesso ao julgado, de interesse especialmente para o pesquisador empírico jurídico, que além de dados estatísticos, geralmente deseja encontrar traços particulares nos julgados pré-selecionados da população estudada.

Por exemplo, este pesquisador pode usar certas ferramentas e técnicas para separar julgados de procedência e examinar características de interesse. É por isso que, "[...] além da integralidade dos acórdãos, o banco de dados deve necessariamente disponibilizar os julgados em seu inteiro teor e permitir que o sistema varra toda a decisão em busca do termo de pesquisa utilizado – e não apenas a ementa ou indexação." (Veçoso *et al.*, 2014, p. 112).

A questão da *pesquisa no inteiro teor* é significativa, pois em geral a pesquisa de jurisprudência feita através de busca textual nos bancos de jurisprudência não é realizada no inteiro teor do acórdão e sim no que é chamado de *espelhos*.

O espelho é um documento que "[...] contém todas as informações contidas no inteiro teor, porém de forma organizada em campos de consulta"

7 O STF explica que atualmente, a atividade de agrupamento é realizada com o suporte de ferramenta tecnológica que, além de verificar o preenchimento dos parâmetros objetivos de vinculação, realiza a comparação entre os textos, a partir de modelo de representação vetorial de palavras (word embedding) construído por meio de redes neurais (neural networks).

8 Há ainda problemas porque os tribunais simplesmente não seguem as recomendações do CNJ. Por exemplo, no Pangea deveriam ser disponibilizados precedentes de todos os tribunais estaduais,8mas, ao acessar a ferramenta, percebe-se que diversos tribunais estaduais não estão incluídos, como o do Acre, Alagoas, Amapá, Ceará, Mato Grosso do Sul, Paraíba, Pernambuco, Piauí, Roraima, Sergipe e Tocantins, o que já é um problema. Além disso, os tribunais eleitorais também não estão integrados no sistema, o que reduz a validade caso seja calculado o tamanho de amostra utilizando-se essa base, que em tese deveria conter todos os tribunais.

(Veçoso *et al.*, 2014, p. 124), criado pela secretaria de cada tribunal com os dados e metadados sendo organizados. Ou seja, a pesquisa feita no banco de jurisprudência, quando feita pelo usuário externo, não percorre todo o inteiro teor, mas somente o espelho, varrendo os campos de pesquisa indicados pelo tribunal respectivo.

Acessar o inteiro teor é importante para uma análise qualitativa de qualidade e até mesmo para a interpretação correta dos dados, na medida em que confiar a pesquisa somente na citação literal do termo não garante confiabilidade: pode se tratar de uma citação para negação da presença do assunto naquele julgado, pode ser uma situação resultado de recursos distintos, pode se tratar de um caso em que se cita o tema na ementa apenas para citar a jurisprudência, mas o julgamento não trata daquele tema específico. A análise do inteiro teor também possibilita a utilização de uma diversidade maior de métodos de pesquisa, a utilização de diversas ferramentas, a exploração do julgado para uma análise mais acurada que incremente a interpretação.

Destarte, cumpre chamar atenção para a necessidade de solução acerca da possibilidade de compartilhamento dos dados estruturados, com o inteiro teor do julgado, ou seja, não adiantaria disponibilizar de acordo com as TPU, pois estas não contêm o inteiro teor. Enquanto esses problemas não são sanados pelas fontes dos dados, o que se pode fazer é seguir uma metodologia possivel de pedidos de compartilhamento dos dados estruturados e registro de todas as situações da coleta de dados.

4 COMPROVANDO ESTATISTICAMENTE PROBLEMAS DOS SISTEMAS ONLINE DE BUSCA DE JURISPRUDÊNCIA ESTADUAIS

Diante da falta de uniformização nos bancos de jurisprudência disponibilizados pelos tribunais do país, problema que traz dificuldades quando o objetivo é calcular o tamanho de amostra ou fazer análises comparativas, acredita-se que o ideal seria que os tribunais utilizassem os mesmos sistemas, a mesma forma de indexação e os mesmos conectores para a pesquisa, além das mesmas possibilidades de pesquisa sobre ementas, inteiro teor, datas de disponibilização do julgado e assim por diante, permitindo que os dados fossem

extraídos de fontes uniformes, valorizando o trabalho com precedentes.

Com intuito de avaliar a qualidade da indexação utilizada nos mecanismos de buscas de jurisprudências dos tribunais estaduais, uma amostra aleatória para o total de jurisprudências dos tribunais foi coletada, considerando a população finita de acórdãos resultantes da busca de jurisprudências seguindo os seguintes critérios: “dolo e improbidade e administrativa”, com filtro de período do julgamento entre 01 de janeiro de 2019 e 31 de dezembro de 2023, presente no Quadro 1 e, posteriormente, foi estratificada proporcionalmente entre os tribunais.

Para fins de cálculo do tamanho mínimo da amostra, consideramos o erro máximo admissível (E) de 5%, confiança de 95% (nível de significância α = 5%) com aproximação da distribuição normal e a variância máxima, ou seja, proporção de processos em não conformidade p = 0,5. Adotando a expressão $n = \frac{N.p.q.(Z_{\alpha/2})^2}{(N-1).E^2+p.q.(Z_{\alpha/2})^2}$ (Agrononik; Hirakata, 2011), com correção para populações finitas, $q = 1 - p$ e N = 19.536 (total de acórdãos resultante da busca em todos os 8 tribunais estaduais apresentados no Quadro 1), temos que $n = \frac{19.536*0,5*(1-0,5)*(1,96)^2}{(19.536-1)*0,05^2+0,5*(1-0,5)*(1,96)^2} = 376,85$. Portanto, o tamanho mínimo amostral necessário será de n = 377 acórdãos. O resultado do processo de estratificação proporcional das amostras de acórdãos pelos tribunais estaduais está apresentado na **Tabela 1**. A amostra final para cada tribunal considerou o primeiro valor inteiro positivo, maior que a amostra proporcional e, arbitrando uma amostra mínima de 20 acórdãos para os tribunais que, proporcionalmente, tiveram uma quantidade menor indicada pela estratificação. Assim, a amostra final do estudo contemplou 413 acórdãos divididos proporcionalmente entre os 8 tribunais estaduais avaliados.

Após o sorteio aleatório dos julgados utilizando o pacote *“sampling”* do RStudio versão 2023.09.1+494 e *seed* = 100, um especialista em direito avaliou o inteiro teor dos processos selecionados e classificou em *conforme* (quando o processo estava de acordo com a indexação e tratava de casos de improbidade administrativa e dolo) ou *não conforme* (quando o processo apenas citava os termos “improbidade administrativa e dolo”, mas não era o assunto principal); ver

Tabela 2.

Tabela 1 - Cálculo amostral

Órgão	N° de acórdãos	%	Amostra estratificada	Amostra final
TJSP	6.151	31,5%	118,7	119
TJMG	3.985	20,4%	76,9	77
TJDFT	3.933	20,1%	75,9	76
TJPR	2.695	13,8%	52,0	52
TJRS	1.270	6,5%	24,5	25
TJPE	731	3,7%	14,1	22
TJRJ	559	2,9%	10,8	21
TJAM	212	1,1%	4,1	21
TOTAL	**19.536**	**100,0%**	**377**	**413**

Fonte: elaboração própria (2024)

Em seguida, intervalos de confiança para a verdadeira proporção de acórdãos indexados em não conformidade (por tribunal) foram. Tais intervalos foram obtidos a partir da seguinte expressão (Magalhães; Lima, 2008, p. 249):

$$IC_{p;(1-\alpha)\%} = \left[\hat{p} - z_{\alpha/2}.\sqrt{\frac{\hat{p}.\hat{q}}{n}}; \hat{p} + z_{\alpha/2}.\sqrt{\frac{\hat{p}.\hat{q}}{n}}\right]$$

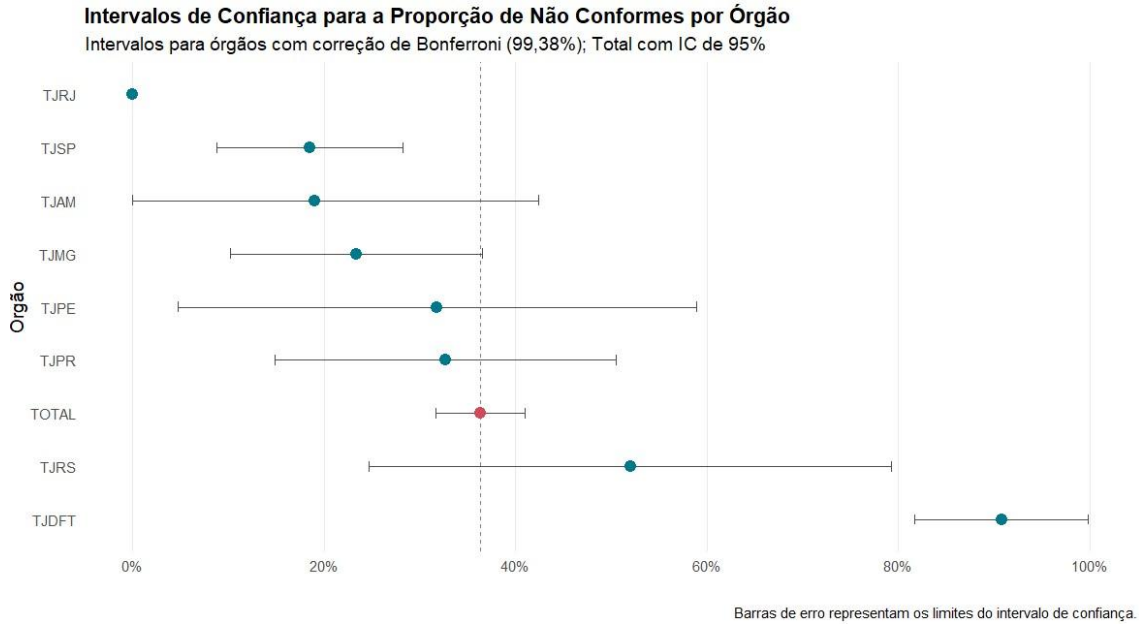
Tabela 2 - Proporção de acórdãos não conformes, por tribunal.

Órgão	n	n° de não conformes	% não conformes	Erro padrão
TJSP	119	22	18,5%	3,6%
TJMG	77	18	23,4%	4,8%
TJDFT	76	69	90,8%	3,3%
TJPR	52	17	32,7%	6,5%
TJRS	25	13	52,0%	10,0%
TJPE	22	7	31,8%	9,9%
TJRJ	21	0	0,0%	0,0%
TJAM	21	4	19,0%	8,6%
TOTAL	**413**	**150**	**36,3%**	**2,4%**

Fonte: elaboração própria (2024)

Para o total acórdãos dos tribunais, a confiança (1-α)% adotada foi de 95%. Ao estratificar por tribunal (m = 8), levamos em consideração a correção de Bonferroni (Lee; Lee, 2018) e o novo nível de confiança ajustado será $(1 - \alpha/m)\% = 99{,}375\%$ para cada um dos intervalos.

Figura 3 - Resultado da amostra de análise dos tribunais estaduais em relação à não conformidade da indexação utilizada nos mecanismos de buscas de jurisprudência



Fonte: elaboração própria (2024).

Desse modo, mostra-se, com 95% de confiança, que a verdadeira proporção de processos indexados em não conformidade considerando o total dos tribunais analisados esteja entre 31,3% e 40,9% (Figura 3). Porém, quando avaliamos por tribunal, temos que o TJRS e o TJPE apresentaram as maiores amplitudes intervalares, ou seja, intervalos menos precisos, com amplitude um pouco superior a 54,0%. O TJDFT tem a maior proporção média de acórdãos relacionados a dolo e improbidade administrativa indexados incorretamente (90,8%). Ainda, com 99,73% de confiança, esperamos que a verdadeira proporção populacional varie entre 81,7% e 99,8%, sendo superior a todos os demais tribunais analisados e considerado o pior cenário. A amostra de 21 acórdãos selecionados aleatoriamente do TJRJ esteve 100% em conformidade com a busca realizada, isto significa dizer que sua indexação de jurisprudências associadas a dolo e improbidade administrativa está sendo feito de maneira adequada. Já o tribunal com maior valor número de acórdãos filtrados, TJSP, tem intervalo de confiança de 99,73% para a proporção de acórdão indexados incorretamente variando entre 8,8% e 28,2%, e proporção média (18,2%) inferior aos demais tribunais.

Em suma, através da Figura 3, vê-se a existência de uma dificuldade em

apenas utilizar palavras chaves para mapear e selecionar julgados, porque muitas vezes os termos escolhidos podem simplesmente ser citados no julgamento que não é sobre aquele tema, contudo o mecanismo de indexação não está calibrado para fazer essa diferenciação, o que resulta em resultados incorretos e na necessidade de triagem manual, o que dificulta ainda mais a confecção de pesquisas, pois imagine-se a triagem manual de centenas, muitas vezes milhares de julgados, para verificar se eles de fato tratam do assunto pesquisado?

Os resultados da análise demonstram a necessidade de pesquisadores considerar que a porcentagem de erro de indexação nos sistemas de busca dos tribunais analisados é considerável. Esse é um dado que deve ser calculado e fazer parte da pesquisa, para cálculo dos níveis de confiabilidade.

5 CONCLUSÃO

A realização de uma pesquisa que contenha os atributos de validade e confiabilidade dependerá diretamente do respeito a uma metodologia de pesquisa bem estruturada, porém flexível para se adaptar aos percalços encontrados pelo pesquisador. O primeiro passo dessa metodologia, conforme exposto no trabalho ora realizado, é uma coleta de dados organizada e registrada e para viabilizar essa fase, o primeiro passo a ser dado pelo pesquisador será entender que o*nde* e *como* coletar dados dependerá diretamente do objetivo e das perguntas da pesquisa. A fase de coleta mantém com esses itens uma relação intrínseca que deve ser constantemente revisada e ao longo da pesquisa é natural que ocorram mutações que irão refletir no desenvolvimento da pesquisa.

Portanto, é importante tentar acessar dados estruturados e, nessa linha, existindo dificuldades na extração direta de dados, o próximo passo poderá ser o pedido de *compartilhamento dos dados*, o que pode ser feito através de requisições enviadas aos tribunais responsáveis. Tais requisições podem se fundamentar na Lei de Acesso à Informação (Lei nº 12.527, de 18 de novembro de 2011) ou em pedidos administrativos baseados nos regimentos de cada ente jurisdicional, tratando-se de pesquisa que exige acesso a dados jurisprudenciais.

É possivel ainda fundamentar o pedido de compartilhamento na Resolução do Conselho Nacional de Justiça 325/2020, que dispõe sobre planejamento estratégico do Poder Judiciário, bem como na Resolução 345 de 2020, que dispõe sobre tecnologia da informação e gestão de demandas judiciais.

É essencial a tentativa de acesso a dados estruturados e somente na ausência dessa possibilidade, deve o pesquisador se voltar ao uso de sistemas de busca online de jurisprudência. Nesse caso, é obrigação do pesquisador ser ainda mais cuidadoso no decorrer da fase de coleta e por isso deverá estar ciente sobre o fato dos sistemas de busca externa não refletirem a completude dos julgados daquele tribunal e por isso não fornecerem um retrato fiel dos julgamentos realizados em determinado período temporal, além de dificultarem pesquisas no inteiro teor do julgado de forma ampla. Essas são situações que podem causar problemas para a confiabilidade e validade das pesquisas realizadas e devem ser abordadas com cuidado.

A falta de uniformização entre os diferentes sistemas de busca online de jurisprudência, que mudam de Estado para Estado, de Região para Região e até do primeiro para o segundo grau de jurisdição também é fato que deve ser objeto de atenção do pesquisador. Certo é que pesquisas valiosas poderiam realizar-se acerca de diversos temas, inclusive congregando análise de dados jurisprudenciais e uma abordagem teórica qualificada, como, por exemplo, a confecção de uma análise comparativa entre julgamentos de diferentes tribunais sobre o mesmo tema[9]. Porém, como fazer isso se cada tribunal tem um sistema de jurisprudência próprio, com conectores distintos, enquanto outro possibilita a pesquisa no inteiro teor e o outro apenas na ementa, um disponibiliza de acordo com a data de publicação, outro com a data de julgamento? Uma consulta aos Tribunais estaduais do Rio de Janeiro e de São Paulo, que podem ser tomados como referências para um estudo mais aprofundado, acerca das técnicas de indexação utilizadas, uma vez que apresentaram menores proporções

[9] Note-se que Kastellec (2010) diz que ”A necessary condition for a legal system to be considered as operating under a rule of law is a high degree of consistency between similar fact situations and similar judicial outcomes: if two courts hear cases with the same facts, the likelihood that both come to the same decision should be high.” E isso seria fundamental para formar uma linha jurisprudencial coerente e organizada.

intervalares de acórdãos indexados em não conformidade com o tema pesquisado.

Sobrestado pelas dificuldades no compartilhamento de dados estruturados, o pesquisador, em regra, é direcionado para a realização da consulta nos sistemas de buscas externos. E assim logo irá se deparar com a falta de uniformização que deve chamar atenção do pesquisador para as *definições conceituais e termos* que irão guiar a coleta de dados, que são diferentes em cada sistema de busca utilizado. Há definições que dependem do viés a ser adotado na pesquisa e da fundamentação teórica que será desenvolvida, sendo recomendável que isso seja exposto de forma clara na metodologia. Alguns guias para ajudar a escolher que definição pode ser buscada é questionar se existem sinônimos, se a ideia é pesquisar a compreensão de uma certa teoria jurídica sobre o termo, se há uma legislação correlacionada àquela definição que utilize termos que podem ser pesquisados em conjunto. O uso de Tesauros é fortemente recomendado, conforme exposto no decorrer desse trabalho.

Deparando com esses problemas, um dos passos mais importantes para o pesquisador será *registrar* integral e rigorosamente todo o processo de coleta de modo documental e ininterrupto. Convém registrar inclusive as tentativas malogradas que possam ter conduzido a mudanças e adaptações na pesquisa, bem como registrar as tentativas bem-sucedidas de diferentes maneiras, pois esse registro auxilia na garantia de confiabilidade e validade das medidas, permitindo a reprodução da pesquisa nas hipóteses de interesse.

É importante, especialmente quando inexistente o acesso a dados estruturados, que o pesquisador considere utilizar a técnica de amostragem, absolutamente essencial para estudos quantitativos jurisprudenciais, na medida em que mais das vezes um determinado tema pode ser objeto de milhões de processos judiciais em curso e com trânsito em julgado, variar em diferentes graus de jurisdição e em diversos tribunais do país. Não existindo acesso a dados estruturados, muitas vezes será necessária a análise singular de cada julgado e para isso a técnica de amostragem se mostra imprescindível, pois facilitará o trabalho do pesquisador.

Já delineado o esforço dos pesquisadores para garantir confiabilidade e validade em pesquisas quantitativas, deve-se ressaltar que a realização desse modelo de pesquisa no Brasil é um processo que depende também dos esforços dos tribunais, que se realmente desejam desenvolver uma cultura de respeito aos precedentes, necessitam viabilizar os meios para a realização de pesquisas empíricas quantitativas. A construção de uma cultura de precedentes será certamente impulsionada e fortalecida pela realização de pesquisas confiáveis. Analisar o direito através do trabalho com dados é uma forma de compreensão que deve ser estimulada, mas para isso é preciso que exista a apropriação fundamentada dos problemas pela comunidade científica e trabalho para a construção de uma solução para problemas que obstaculizam o intento daqueles que pretendam pesquisar fenômenos de forma quantitativa.

Os obstáculos ora relatados existem e devem ser discutidos, porém podem ser minorados ou ao menos esclarecidos através do cuidado dos pesquisadores na realização de sua pesquisa, pois, enquanto as fontes dos dados não implementam soluções, caberá aos pesquisadores trabalhar com uma metodologia de pesquisa que viabilize a confiabilidade e validade mesmo quando uma pesquisa com dados jurisprudenciais tenha que recorrer a sistemas de busca online de jurisprudência.

Espera-se que brevemente pesquisar empiricamente o direito dependa não da superação de impedimentos, mas do desejo de alunos e pesquisadores de se voltar para o uso da estatística para auxiliar o sistema jurídico nessa nova construção de conhecimento, um campo novo e rico a ser explorado.

REFERÊNCIAS

AGRANONIK, M.; HIRAKATA, V. N. Cálculo de tamanho de amostra: proporções. **Clinical and Biomedical Research,** Porto Alegre: v. 30, n. 31, p. 382-388, 2011

BRASIL. SUPERIOR TRIBUNAL DE JUSTIÇA. **STJ - Vocabulário Jurídico**, 2024. Disponível em https://scon.stj.jus.br/SCON/servlet/ThesMain?action=consultar&pesquisa=improbidade. Acesso em: 10 jun. 2024.

CONSELHO NACIONAL DE JUSTIÇA. **Pesquisa empírica:** Introdução à pesquisa judiciária**.** Brasília. 2023. Disponível em: https://www.cnj.jus.br/wp-content/uploads/2023/02/apresentacao-como-fazer-pesquisas-empiricas-introducao-pesquisa-judiciaria-23-03-2023.pdf. Acesso em: 12 abr. 2024.

CONSELHO NACIONAL DE JUSTIÇA. **Banco Nacional de Dados de Demandas Repetitivas e Precedentes Obrigatórios**. Disponível em: https://bnp.pdpj.jus.br/. Acesso em: dez. 2025

EPSTEIN, L.; KING, G. **Pesquisa empírica em direito:** As regras de inferência. São Paulo: Direito GV, 2013. Disponível em https://repositorio.fgv.br/server/api/core/bitstreams/963518b6-c0ab-4cf7-acc1-a5aa2b2f84ea/content. Acesso em: fev. 2024.

KASTELLEC, J. P. Trees and regression in empirical legal studies. **Journal of Empirical Legal Studies**, [*S.l.*], v. 7, n. 2, p. 202-230, 2010. Disponível em: https://jkastellec.scholar.princeton.edu/sites/g/files/toruqf3871/files/jkastellec/files/trees.pdf. Acesso em: 09 jun. 2024.

LEE, S; LEE DK. What is the proper way to apply the multiple comparison test? **Korean J Anesthesiol**, [*S.l.*], n. 71, v. 5, p. 353-360, 2018.

MAGALHÃES, M. N; LIMA, A. C. P. **Noções de Probabilidade e Estatística**. 6 ed. São Paulo: Edusp, 2008.

PINHEIRO, L. V. R.; FERREZ, H. D. **Tesauro Brasileiro de Ciência da Informação**. Rio de Janeiro; Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), 2014.

PROGRAMA DE GESTÃO DE PRECEDENTES DO PODER JUDICIÁRIO. **Pangea/BNP – Precedentes Qualificados.** Disponível em: https://pdpj.jus.br. Acesso em: 10 jun. 2024.

TRECENTI, J. A. Z. **Diagramas de influência:** uma aplicação em Jurimetria. 2016. Dissertação (Mestrado) – Universidade de São Paulo, São Paulo, 2016. Disponível em: https://teses.usp.br/teses/disponiveis/45/45133/tde-20230727-113325/. Acesso em: 04 jun. 2024.

VEÇOSO, F. F. C., PEREIRA, B. R., PERRUSO, C. A., MARINHO, C. M., BABINSKI, D. B. O., WANG, D. W. L., GUERRINI, E. W., PALMA, J. B., & SALINAS, N. S. C. A pesquisa em direito e as bases eletrônicas de julgados dos tribunais: matrizes de análise e aplicação no Supremo Tribunal Federal e no Superior Tribunal de Justiça. **Revista de Estudos Empíricos em Direito/Brazilian Journal of Empirical Legal Studies**, [*S.l.*], n. 1, v. 1, p. 105-139, 2014.

THE RELIABILITY OF JURIMETRICS RESEARCH: CHALLENGES IN DATA COLLECTION AND EMPIRICAL CASE LAW ANALYSES

ABSTRACT
Objective: To underscore the necessity of a rigorous and specific methodology for validating the reliability of quantitative research within the legal field, while demonstrating how deficiencies in the data collection phase may compromise both the validity and the reliability of such research. **Methodology:** Exploratory analysis with qualitative and quantitative approaches. The research was conducted by examining theoretical works that stipulate how jurimetric research should be carried out, and normative publications from the National Council of Justice and the Supreme Federal Court that regulate the subject. Following this, external case law systems of eight state courts were examined. **Results:** The article demonstrates how these problems significantly hinder or even make impossible the quantitative empirical research of case law and statistically exposes the quality of indexing used in the search engines of case law from the State Courts of São Paulo, Rio Grande do Sul, Paraná, Pernambuco, Minas Gerais, the Federal District, Rio de Janeiro, and Amazonas. **Conclusions:** The quantitative analysis revealed the problems arising from the poor indexing quality and lack of standardization across eight external Brazilian case law search databases, and methods were proposed to overcome these difficulties.

Descriptors: Data Collection. Data Analysis. Legal Databases; Jurimetrics; Standardization.

LA CONFIABILIDAD EN LA INVESTIGACIÓN JURIMÉTRICA: DESAFÍOS EN LA RECOLECCIÓN DE DATOS AND ANÁLISIS EMPÍRICOS DE JURISPRUDENCIA

RESUMEN
Objetivo: Evidenciar la necesidad de una metodología específica y rigurosa para validar la fiabilidad de las investigaciones cuantitativas en el ámbito jurídico, demostrando cómo los problemas en la fase de recolección de datos afectan negativamente la validez y la fiabilidad de dichas investigaciones. **Metodología:** Análisis exploratorio con enfoque cualitativo y cuantitativo. La investigación se llevó a cabo mediante el examen de trabajos teóricos que estipulan cómo debe realizarse una investigación de jurimetría y publicaciones normativas del Consejo Nacional de Justicia y del Tribunal Supremo Federal que regulan el tema, para posteriormente examinar sistemas externos de jurisprudencia de ocho tribunales estatales. **Resultados:** Se demuestra cómo estos problemas dificultan enormemente o incluso hacen imposible la investigación empírica cuantitativa de jurisprudencia y se expone estadísticamente la calidad de la indexación utilizada en los motores de búsqueda de jurisprudencia de los Tribunales Estatales de São Paulo, Rio Grande do Sul, Paraná, Pernambuco, Minas Gerais, Distrito Federal, Rio de Janeiro y Amazonas. **Conclusiones:** El análisis cuantitativo mostró los problemas derivados de la mala calidad de la indexación y la falta de uniformidad en ocho bases de datos externas de búsqueda de jurisprudencia brasileñas, y se propusieron métodos

para superar estas dificultades.

Descriptores: Análisis de Datos. Recopilación de Datos. Bancos de Jurisprudencia. Jurimetría. Uniformización.

Recebido em: 25.02.2025
Aceito em: 30.09.2025