UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Yuri Alexandre dos Santos

Ciência de dados aplicada à jurimetria: um estudo de caso a partir de decisões do Tribunal de Justiça de Minas Gerais

Belo Horizonte
2024

Yuri Alexandre dos Santos

Ciência de dados aplicada à jurimetria: um estudo de caso a partir de decisões do Tribunal de Justiça de Minas Gerais

Versão final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Wagner Meira Júnior

Belo Horizonte
2024

Santos, Yuri Alexandre dos.

S237c Ciência de dados aplicada à jurimetria: [recurso eletrônico] um estudo de caso a partir de decisões do Tribunal de Justiça e Minas Gerais / Yuri Alexandre dos Santos - 2024.

1 recurso online (55 f. il., color.) : pdf.

Orientador: Wagner Meira Júnior.

Dissertação (Mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 52-55

1. Computação – Teses. 2. Ciência de dados – Teses. 3. Jurimetria – Teses. 4. Sentenças judiciais – Minas Gerais - Teses. 4. Defesa do consumidor – Danos morais – Teses. I. Meira Júnior, Wagner. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III. Título.

CDU 519.6*41(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

CIÊNCIA DE DADOS APLICADA À JURIMETRIA: UM ESTUDO DE CASO A PARTIR DE DECISÕES DO TRIBUNAL DE JUSTIÇA DE MINAS GERAIS

YURI ALEXANDRE DOS SANTOS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Wagner Meira Júnior - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Leonardo Netto Parentoni
Faculdade de Direito - UFMG

Prof. Virgílio Augusto Fernandes Almeida
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 25 de março de 2024.

Documento assinado eletronicamente por **Wagner Meira Junior**, **Professor do Magistério Superior**, em 28/03/2024, às 11:30, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Virgilio Augusto Fernandes Almeida**, **Professor Magistério Superior - Voluntário**, em 01/04/2024, às 12:17, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Netto Parentoni**, **Professor do Magistério Superior**, em 05/06/2024, às 15:21, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3138542** e o código CRC **0686F0F4**.

Resumo

A jurimetria pode ser definida como o uso de métodos quantitativos aplicáveis ao mundo jurídico. Dentro da recente tendência que se vem verificando no Direito, não sem forte oposição, no sentido de estudá-lo sob um viés mais empírico do que propriamente filosófico ou exclusivamente jurídico, este trabalho busca propor uma metodologia de análise de decisões judiciais a partir da utilização de técnicas de jurimetria em dados não estruturados. Os dados utilizados para análise e construção da referida metodologia consistem de sentenças judiciais proferidas em primeiro grau, no âmbito do Tribunal de Justiça de Minas Gerais, em demandas cuja causa de pedir principal é a condenação da parte ré a indenizar a parte autora em virtude de ter inscrito seu nome indevidamente em cadastros de inadimplentes. Espera-se, em última análise, que este trabalho possa contribuir para o avanço da jurimetria no Brasil.

Palavras-chave: jurimetria; análise automatizada de sentenças judiciais; indenização por danos morais.

Abstract

Jurimetrics may be defined as the usage of quantitative methods to study and understand the legal field. Considering the recent trend verified in Law to study it under a more empirical, rather than a philosophical or an exclusively legal, approach, this work proposes a pipeline for analyzing judicial rulings through the usage of jurimetrics techniques in non-structured data. The data used for this study consists in judicial rulings issued by judges from the Court of Justice of Minas Gerais, which is one of the biggest in Brazil, in legal claims in which the plaintiff sues the defendant for pain and suffering caused from being unlawfully included in credit restraint lists. This work aims at contributing to the advancement of jurimetrics in Brazil.

Keywords: jurimetrics; automated analysis of judicial rullings; pain and suffering damages.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Contexto

Segundo o relatório "Justiça em Números 2021" do Conselho Nacional de Justiça, o número de novos processos somente no âmbito do Tribunal de Justiça de Minas Gerais (TJMG) foi de 1.428.480, sendo que o tribunal já possuía, naquele ano, 3.940.277 casos pendentes (Brasil, 2021, p. 47). A taxa total de congestionamento daquele tribunal (correspondente ao "percentual de processos que ficaram represados sem solução, comparativamente ao total tramitado no período de um ano" (Brasil, 2021, p. 127) foi de 70,8%, e a taxa líquida de congestionamento (equivalente à total, mas desconsiderando "processos suspensos, sobrestados ou em arquivo provisório" (Brasil, 2021, p. 132) de 72,7%.

No Brasil, considerando todos os tribunais de justiça estaduais, as taxas ficaram em 72,5% e 75,0% (Brasil, 2021, p. 132), respectivamente, dado que demonstra a dificuldade que os tribunais têm para lidar com o alto volume de demandas distribuídas todos os anos. Com efeito, o tempo médio de giro do acervo de processos (indicador que mostra quanto tempo seria necessário para que magistrados e servidores zerassem o estoque de casos pendentes, mantidas as suas respectivas produtividades (Brasil, 2021, p. 105) na justiça estadual é de três anos, caindo para dois anos e oito meses no Tribunal de Justiça de Minas Gerais, exatamente igual ao indicador de todo o Poder Judiciário (Brasil, 2021, p. 108).

Por outro lado, o tempo médio de tramitação de processos que se encontravam pendentes de baixa até 31 de dezembro de 2020 no âmbito do TJMG, foi de 2 anos e 9 meses no primeiro grau e de 3 anos e 6 meses no segundo (Brasil, 2021, p. 209-211).

Como se nota, e como já há muito se sabe entre os que atuam no dia-a-dia da prática forense, o desafio do Poder Judiciário nacional é grande no que diz respeito à garantia razoável duração do processo, prevista no artigo 5º, inciso LXXVIII, da Constituição Federal (Brasil, 1988). Com efeito, o problema da morosidade não é novo (Schwengber, 2006; Britto et al., 2018).

Entre outros fatores, contribui para tal dificuldade o massivo número de demandas em tramitação nos tribunais brasileiros, de modo que iniciativas que busquem contribuir para que os membros do Poder Judiciário possam exercer suas atividades de maneira mais assertiva e eficiente são absolutamente necessárias.

O campo do Direito, seja na academia, seja na prática, ainda se beneficia pouco de estudos e de levantamentos empíricos, notadamente daqueles baseados na análise e na modelagem quantitativas de dados, ainda que a necessidade de se realizarem estudos do tipo já venha sendo apontada há algum tempo no Brasil e no exterior.

Este estudo pretende enfrentar, assim, um duplo desafio: do lado do Direito, busca realizar um estudo quantitativo com dados jurídicos, sendo esta abordagem ainda pouco frequente no cenário nacional; e, no campo da Computação, busca criar um método de coleta e análise e modelagem de dados não estruturados oriundos de processos judiciais, os quais ainda são pouco explorados no Brasil.

1.1.1 Breves considerações sobre o processo judicial

Um processo judicial pode ser definido como "um procedimento em contraditório destinado à construção dos provimentos estatais, em que todos os sujeitos interessados participam, em igualdade de

condições, na produção do resultado" (Alexandre Freitas Câmara, 2022) e é comumente dividido em duas etapas: uma denominada *cognitiva* (também conhecida como fase ou processo de *de conhecimento*) e outra *executiva* (também chamada de fase ou processo *de cumprimento de sentença* ou *de execução*). O principal provimento estatal buscado em um processo judicial é a "sentença", definida como "o ato que extingue o processo ou alguma de suas fases" (cognitiva ou executiva)" (Alexandre Freitas Câmara, 2022).

A fase de conhecimento é aquela na qual o juiz decidirá se o autor de fato possui o direito por ele alegado, ou se quem tem razão, por assim dizer, é o réu. Se for possível definir de quem é o direito, o juiz *resolverá o mérito* e proferirá uma sentença definitiva; se não for possível, porém, a sentença será apenas terminativa,[1] isto é, colocará fim à fase de conhecimento, mas não dirá se o direito está com o autor ou com o réu. Assim, na definição de Alexandre Freitas Câmara (2022), a fase (ou o processo) de conhecimento "é o processo de sentença, isto é, o processo que tem por objeto imediato a produção de uma sentença de mérito, declaratória da existência ou inexistência de um direito."

Este trabalho tem seu foco justamente na sentença proferida na etapa cognitiva do processo judicial cível, motivo pelo qual não se entrará em detalhes quanto à definição e quanto aos atos englobados pela fase executiva.

Além disso, é importante saber que as regras que regem o processo judicial podem variar. O processo seguirá o *rito do procedimento comum* se estiver apenas submetido às regras do Código de Processo Civil aplicáveis aos processos em geral, e seguirá outros ritos se estiver submetido a outras normas procedimentais, sendo que, nestes casos, as regras que disciplinam o procedimento comum são aplicáveis em caráter subsidiário. Entre os ritos especiais de tramitação processual, existe aquele adotado no âmbito dos juizados especiais cíveis, regidos pela Lei Federal nº 9.099/1995. Trata-se de um procedimento simplificado, no qual as regras são simplificadas para que o processo seja mais simples e rápido.

Apesar de ser, em geral, mais lento, o procedimento comum permite que se discutam causas de maior complexidade, que precisam de mais tempo para produção de provas: assim, por exemplo, este procedimento permite a realização de perícias técnicas e a oitiva de um número maior de testemunhas. No entanto, para mover uma demanda sob o rito do procedimento comum, em geral a parte autora precisa pagar as custas do processo logo no início, a menos que não tenha condições financeiras de arcar com elas. De qualquer modo, ao final, as custas devem ser pagas pela parte que perder a demanda, assim como um valor devido ao advogado da parte vencedora (denominado "honorários de sucumbência"), calculado como uma porcentagem sobre o valor da causa.

No caso dos juizados especiais cíveis, não há pagamento de custas ou de honorários de sucumbência por nenhuma das partes se o processo se resolver na primeira instância - sem que haja recurso. Isto significa que tais processos, além de geralmente mais rápidos, também tendem a ser mais baratos - muito embora com a simplicidade venha a impossibilidade de produzir provas complexas.

1.1.2 Problema abordado e contribuições

O presente trabalho se ocupa de demandas judiciais relacionadas ao Direito do Consumidor, mais especificamente aos casos em que o autor alega ter sofrido dano moral em virtude de seu nome ter sido indevidamente inscrito em cadastros de inadimplentes. Se observados os requisitos da Lei Federal nº 9.099/1995 - e geralmente os requisitos se fazem presentes em casos como os aqui tratados -, este tipo de demanda pode ser ajuizado, a critério do autor,[2] na justiça comum, na qual o processo seguirá o rito do procedimento comum, ou em um juizado especial cível, no qual seguirá o rito simplificado já mencionado anteriormente. Em se tratando de uma escolha do autor, é possível que a opção por um ou outro foro interfira no resultado da demanda, sendo esta uma das investigações que este estudo pretende realizar.

Entre os processos que ingressaram no Poder Judiciário brasileiro a nível estadual em 2020, o assunto "DIREITO DO CONSUMIDOR - Responsabilidade do Fornecedor/Indenização por Dano Moral"

[1] Segundo Alexandre Freitas Câmara (2022), "terminativa é a sentença que não contém a resolução do mérito da causa; definitiva, a que contém a resolução do mérito."

[2] Há quem defenda, inclusive no âmbito do Poder Judiciário, que o foro dos juizados especiais cíveis é obrigatório se a demanda atender aos requisitos definidos na Lei dos Juizados Especiais. No entanto, tal interpretação não encontra respaldo no texto da lei, que parece ser muito claro no sentido de que o foro é facultativo, motivo pelo qual aqui se adota a interpretação de que o autor pode definir se ajuizará uma demanda como as que são objeto deste estudo segundo o procedimento comum ou segundo aquele previsto para os juizados especiais cíveis.

foi o quarto mais demandado no segundo grau (com 253.410 processos, representando 1,90% do total), o primeiro mais demandado nas turmas recursais (com 254.155 processos, correspondentes a 12,88% do total) e o primeiro mais demandado nos juizado especiais cíveis (com 635.296 processos, 8,87% do total) (Brasil, 2021, p. 275-276).

O dano moral corresponde àquele que atinge um direito de personalidade. Nas palavras de Pereira, "o fundamento da reparabilidade pelo dano moral está em que, a par do patrimônio em sentido técnico, o indivíduo é titular de direitos integrantes de sua personalidade, não podendo conformar-se a ordem jurídica em que sejam impunemente atingidos" (Pereira, 2022).

O assunto é tratado, pelo Código Civil Brasileiro, da seguinte maneira: em seu artigo 186, a lei dispõe que a pessoa física ou jurídica que causa a outra um dano moral (ou seja, um dano a algum de seus direitos de personalidade) pratica ato ilícito, sendo que o causador do dano, por força do disposto em seu artigo 927, deve repará-lo.

Acontece, porém, que, se por um lado não há dúvidas de que o dano moral gera responsabilidade civil, por outro não há parâmetros legais suficientemente claros e específicos para a fixação de um valor que seja apto a reparar um determinado dano a direito de personalidade. A questão é complexa, na medida em que as métricas utilizadas para a quantificação de danos materiais não se aplicam aos danos morais (com efeito, não é possível dizer o quanto "vale" a honra ou a imagem de alguém, por exemplo).

À falta de lei, a fixação de critérios para a reparação de danos morais cabe à doutrina e à jurisprudência. Quando do julgamento do Recurso Especial 1.374.284/MG, afetado ao rito dos recursos repetitivos, o Superior Tribunal de Justiça definiu critérios para fixação de reparação por danos morais, quais sejam: o grau de culpa do infrator, o nível socioeconômico do autor e o porte do infrator. Além disso, o juiz deve orientar-se "pelos critérios sugeridos pela doutrina e jurisprudência, com razoabilidade, valendo-se de sua experiência e bom senso, atento à realidade da vida e às peculiaridades de cada caso, de modo a que, de um lado, não haja enriquecimento sem causa de quem recebe a indenização e, de outro, haja efetiva compensação pelos danos morais experimentados por aquele que fora lesado" (Brasil, 2014).

Como se nota, em que pese o louvável esforço daquele tribunal no sentido de definir critérios para a fixação de reparação por danos morais, fato é que os parâmetros por ele fixados seguem abstratos (merecendo destaque a utilização do termo "sugeridos" no excerto supracitado). Na prática, considerando o também elevado grau de abstração dos parâmetros comumente definidos pela doutrina nacional, cabe ao juiz, no caso concreto, arbitrar os valores que entender mais adequados à reparação do dano que for levado à sua apreciação.

Como não há critérios bem definidos, acredita-se que o estudo da forma como magistrados fixam danos morais em casos concretos pode contribuir para uma maior efetividade na prestação jurisdicional. Do lado do Poder Judiciário, o conhecimento obtido a partir de um levantamento como esse tem o condão de levar a uma padronização de práticas em casos similares, inclusive com a eventual redução do número de recursos (ou mesmo de processos, já que um jurisdicionado pode deixar de mover uma demanda meramente "aventureira", que lhe renderá pouco ou nenhum rendimento). Do lado das partes, o acesso a informações que as auxiliem a compreender suas reais chances no âmbito de uma determinada demanda pode favorecer a autocomposição ou ao menos prevenir a propositura ou a continuidade de demandas cujos retornos financeiros sejam baixos ou irrisórios.

No caso específico deste estudo, considerando que a escolha entre mover uma ação de indenização por danos morais segundo o rito ordinário, na justiça comum, ou segundo o rito dos juizados cabe ao postulante, aferir a eventual existência de padrões decisórios distintos em uma ou outra via pode contribuir na decisão entre um ou outro juízo. Por exemplo: se confirmado o senso comum que existe entre os que atuam no cotidiano forense de que os valores de indenização praticados na justiça comum são superiores àqueles fixados nos juizados especiais, um postulante poderia optar pela primeira via se seu objetivo fosse maximizar seu ganho financeiro.

Como o dano moral pode surgir no âmbito dos mais diversos tipos de relações, este trabalho teve de fazer um recorte no assunto, sob pena de se tornar inexequível ou excessivamente complexo. Assim, optou-se por estudar um tipo específico de demanda: a que versa sobre o dano moral gerado a alguém em virtude de sua inscrição indevida em cadastros de inadimplentes.

A opção por esta abordagem tem duas vantagens principais.

Em primeiro lugar, é importante considerar que este tipo de demanda é extremamente recorrente no Judiciário mineiro. Para demonstrá-lo, basta ressaltar que o Superior Tribunal de Justiça mapeou a existência de vinte teses sobre o tema "cadastro de inadimplentes" oriundas somente daquele tribunal, em julgados publicados até 29 de abril de 2016 (Brasil, 2016). Entre as vinte teses, nove foram julgadas sob o rito dos recursos repetitivos, previsto no artigo 543-C do CPC/73, então vigente (Brasil, 1973).

Em segundo lugar, um dos principais entendimentos do STJ sobre o assunto é no sentido de que "a inscrição indevida em cadastro de inadimplentes configura dano moral *in re ipsa*"(Brasil, 2016, p. 1). Isto é revelante porque, em geral, o dano passível de indenização (no caso dos danos materiais) ou de reparação (no caso dos danos morais) deve ser provado, sendo a prova, em se tratando de ofensa a direitos de personalidade, bastante difícil e, muitas vezes, complexa, já que o sofrimento de cada um é bastante subjetivo.

Processos que tratam de negativação indevida são, em geral, mais simples, na medida em que, para a fixação do dever de reparar o dano ao infrator, basta que se demonstrem: (a) que a inscrição do nome da vítima efetivamente ocorreu e (b) que a inscrição era indevida. Dessa forma, ainda que o recorte escolhido seja restrito, acredita-se que esta baixa complexidade aliada ao elevado número de processos sobre o assunto criam um cenário favorável para a utilização de métodos quantitativos no processamento e na análise de dados a ele relacionados.

O presente trabalho realizará um estudo quantitativo com dados não estruturados de processos judiciais julgados em primeira instância no âmbito do Tribunal de Justiça de Minas Gerais, quais sejam, sentenças proferidas na justiça comum e nos juizados especiais cíveis em demandas cujo assunto é a inscrição indevida de pessoas físicas em cadastros de inadimplentes. Esta pesquisa, portanto, se insere no campo da *jurimetria* (termo cujo significado será melhor abordado na próxima seção) e pretende contribuir para o seu avanço no Brasil.[3]

1.2 Uma definição de jurimetria

O termo *jurimetria* foi cunhado por Loevinger (Baade, 1963, p. 1) e, em sua primeira acepção, faz referência à "investigação científica de problemas jurídicos" (p. 483, Loevinger, 1949, tradução nossa).

O autor, em seu texto de 1949, aponta o problema de que, já naquela época, o direito e aqueles que o manipulam padeciam da dificuldade de se fazerem entender. Ao longo de sua argumentação, critica séculos de pensamento e prática jurídicos destinados à discussão de abstrações com pouca utilidade prática. O autor critica, em especial, o papel da jurisprudência na sedimentação desta tendência, afirmando que esta última vem "se concentrando em responder questões sem sentido por meio de especulação fútil" (Loevinger, 1949, p. 470).

Para Loevinger, o estudo e a aplicação do Direito eram, à sua época, feitos de forma especulativa, sem rigor científico, fato que ele diz ser comum às ciências sociais, em oposição às ciências exatas e às ciências biológicas. Segundo ele, "o Direito não ainda não estabeleceu técnicas institucionais para produzir conhecimento científico sobre seus problemas" (p. 476, Loevinger, 1949, tradução nossa).

Assim, para resolver o problema da falta de uma abordagem científica no Direito, (p. 483, Loevinger, 1949, tradução nossa) afirma que "o próximo passo no longo caminho do progresso humano deve ser da jurisprudência (que é uma mera especulação sobre o Direito) para a jurimetria". Para ele, portanto, a jurimetria é a evolução da jurisprudência, mais qualificada na medida em que busca na ciência seus métodos e ferramentas para compreender e resolver os problemas do Direito.

Ao final de sua exposição, Loevinger afirma não ser possível, sem o avanço dos estudos na área, dizer exatamente quais problemas serão abordados pela jurimetria (Loevinger, 1949, p. 484), mas se arrisca a supor alguns. Ao todo, faz nove suposições, entre as quais, para os fins deste trabalho, as de maior interesse são a segunda ("the behavior of judges"), a quarta ("legal language and communication") e a quinta ("legal procedure and recordation") (Loevinger, 1949, p. 485-487), tendo em vista que tais searas são todas, em alguma medida, abordadas nesta análise.

Embora Loevinger tenha formulado suas críticas ao Direito e aos juristas, além destas perguntas, ainda na década de 1940, seus apontamentos permanecem verdadeiramente atuais e se aplicam, em larga medida, ao contexto brasileiro, conforme se verá mais adiante.

[3] A realização de estudos jurimétricos com dados judiciais vem avançando em ritmo lento no Brasil e sofre com empecilhos que vão desde a inexistência de mecanismos facilitados e padronizados de acesso a dados (como *APIs*) ao uso de CAPTCHA| em sistemas de busca processual (Colombo et al., 2017, p. 12). Possivelmente por este motivo, estudos como este são pouco frequentes, e os que existem dependem de acordos com instituições provedoras de dados e do uso de técnicas como coleta manual ou *crawling*.

Seguindo nas ideias expostas em seu escrito inaugural sobre o tema, Loevinger afirma que "é desnecessário, senão impossível, dar uma definição precisa do campo da jurimetria", tarefa que cabe àqueles que nele atuam (Loevinger, 1963, p. 8). Apesar disso, afirma que a área compreende, entre outros, "o uso de lógica matemática no direito" e "a recuperação de dados jurídicos por meios eletrônicos e mecânicos" (Loevinger, 1963, p. 8), pontos estes que serão objeto do presente trabalho.

Com o tempo, o termo "jurimetria", ao menos em âmbito nacional, foi se distanciando das primeiras formulações de Loevinger para se aproximar de uma definição que o vincula ao "uso de métodos estatísticos aplicados à Ciência Jurídica" (Maia & Bezerra, 2020, p. 6). No mesmo sentido caminha a definição de Nunes (2016b).

A definição corrente, porém, tem o inconveniente de ser pouco clara, sobretudo considerando que a expressão "métodos estatísticos" pode englobar ou deixar de englobar uma série de técnicas que, embora se valham de raciocínios e de recursos próprios à Estatística, ficam à margem do que comumente se entende, sobretudo no mundo do Direito, por "método estatístico". É o caso, por exemplo, do aprendizado de máquina, que recorre a conhecimentos daquela área do conhecimento para o enfrentamento de problemas que se propõe a resolver, mas que com ela não se confunde.

Com efeito, o objetivo do aprendizado de máquina é a realização de predições, conquanto a finalidade da estatística é descrever relações (Bzdok et al., 2018). Acontece, no entanto, que o primeiro conceito, tanto quanto o segundo, parece se amoldar ao que Loevinger idealizou como objeto da jurimetria, o que não acontece na definição que se vem adotando no Brasil.

Além disso, a utilização da expressão "Ciência Jurídica" é feita como se a ideia de Direito enquanto ciência fosse dada, pressuposta, à revelia, portanto, dos séculos de estudos dedicados a discutir a natureza do Direito e de suas indagações. Ao que parece, aliás, o Direito é antes tecnologia que ciência, na linha do que defende Moraes (2017).

Assim, não havendo certezas quanto ao que seria um "método estatístico aplicável à Ciência Jurídica" e diante da recusa de Loevinger de apresentar um conceito formal de jurimetria, este trabalho adotará uma definição própria, apta a acomodar as ideias daquele autor. Assim, aqui, *jurimetria* será **o uso de métodos quantitativos aplicáveis ao mundo jurídico**, conceito que nos parece mais adequado em virtude dos fatos de que:

- A expressão "métodos quantitativos" abrange qualquer abordagem que envolva a análise, a modelagem e a utilização de dados sob uma perspectiva quantitativa, incluindo tanto as aplicações puramente estatísticas quanto as que se valem, por exemplo, de técnicas de aprendizado de máquina e de mineração de dados;
- "Mundo jurídico" engloba tudo que se relaciona à criação, à interpretação e à aplicação do direito, este último entendido como um conjunto de normas.

Assim, na medida em que este trabalho, como se verá adiante, pressupõe o uso de técnicas de análise e modelagem de dados, eminentemente estatísticas, e de técnicas de processamento de linguagem natural, com o uso de modelos de aprendizado de máquina, tem-se que suas indagações se amoldam perfeitamente ao conceito aqui proposto, não havendo dúvidas, portanto, de que se enquadra no campo da jurimetria, tal qual inicialmente idealizada por Loevinger.

1.3 Objetivos e hipóteses de pesquisa

O objetivo geral deste trabalho é contribuir para a compreensão do funcionamento do Poder Judiciário sob o prisma da jurimetria, valendo-se de métodos quantitativos oriundos da Ciência da Computação e da Estatística, notadamente nos campos da Ciência de Dados, do Processamento de Linguagem Natural e do Aprendizado de Máquina. Para tanto, propõe uma metodologia para análise, modelagem e comparação de dados não estruturados, quais sejam, arquivos de texto que contêm, cada um, uma sentença proferida por uma vara cível ou por uma vara de juizado especial cível do Tribunal de Justiça de Minas Gerais no âmbito de processo cujo assunto principal é a inscrição do nome de seu respectivo autor em cadastros de inadimplentes.

Seus objetivos específicos, por outro lado, são os seguintes:

1. Construir uma base de dados a partir de informações sobre processos judiciais recebidas do Tribunal de Justiça de Minas Gerais, apta a ser utilizada para fins de caracterização e de modelagem;
2. Caracterizar quantitativamente as sentenças quanto aos seus diferentes aspectos, buscando compreender os fatores que influenciam na fixação de um determinado valor a título de danos morais;
3. Analisar as sentenças judiciais com o emprego de técnicas de aprendizado de máquina;
4. Formalizar, ao final, uma metodologia de análise e modelagem de sentenças que possa ser considerada em estudos futuros, com dados não contemplados por esta pesquisa.

Para que tais objetivos sejam alcançados, formulam-se as seguintes hipóteses, as quais foram objeto de investigação:

1. É possível identificar um padrão bem estabelecido para a fixação de reparações por danos morais;
2. Sentenças fixadas na justiça comum têm valor maior, no geral, que sentenças fixadas no âmbito dos juizados especiais;
3. É possível utilizar modelos de aprendizado de máquina a partir de sentenças judiciais e dos valores por elas fixados a título de indenização para fins de compreendê-los e de fazer inferências a partir deles, com o objetivo de compreender os critérios e os padrões utilizados pelos juízes para definição dos referidos valores;
4. É possível construir um modelo de aprendizado de máquina que seja capaz de, dado o texto corrido de uma sentença no âmbito da qual se fixou reparação por danos morais, identificar o valor da condenação.

Todas as hipóteses foram avaliadas para as sentenças constantes de uma base de dados específica, conforme se trata em maiores detalhes no Capítulo 3 deste trabalho.

1.4 Organização do trabalho

Esta dissertação é organizada em sete capítulos.

Neste primeiro capítulo, foram apresentados o contexto que motiva a pesquisa realizada, o problema abordado, as contribuições que o estudo pretende alcançar e os objetivos geral e específicos da pesquisa. Além disso, delimitou-se o conceito de "jurimetria" aqui adotado, contrapondo-o às concepções comumente adotadas na literatura jurídica brasileira.

O Capítulo 2 faz uma revisão de literatura, buscando, no contexto nacional, entender como a jurimetria tem sido aplicada na prática. Em específico, buscou-se levantar estudos quantitativos realizados com dados atinentes ao mundo jurídico marcados pelo emprego de técnicas computacionais.

Os Capítulos 3, 4 e 5 tratam do método e apresentam os resultados obtidos com a pesquisa: o Capítulo 3 apresenta o que se fez para coletar e consolidar os dados utilizados nas fases seguintes; o Capítulo 4 demonstra os procedimentos utilizados para caracterização dos dados, incluindo uma análise estatística mais clássica; e o Capítulo 5 trata da utilização de técnicas de aprendizado de máquina para análise de sentenças (mais especificamente, de agrupamento e de regressão) e apresenta os resultados obtidos com o desenvolvimento de um classificador destinado à identificação dos valores fixados a título de indenização por danos morais em sentenças judiciais.

O Capítulo 6 discute as análises e resultados obtidos, aponta as limitações da presente investigação e apresenta futuros possíveis caminhos de pesquisa.

Capítulo 2

Revisão de literatura

A jurimetria vem, nos últimos anos, crescendo no meio acadêmico (Maia & Bezerra, 2020, p. 20). Apesar disso, boa parte (senão a maior) dos trabalhos na área, ao menos no Brasil, é composta por escritos de caráter filosófico-jurídico que se limitam a discutir as diversas aplicações da jurimetria, na medida em que não utilizam métodos quantitativos para a realização de suas análises. Tratam-se, antes, de textos mais ligados a outras áreas do Direito, como a Filosofia do Direito.

São exemplos de estudos que buscam investigar a aplicabilidade, as potencialidades e os problemas da utilização da jurimetria a uma determinada situação ou área do Direito os elaborados por Novaes et al. (2018), Zabala & Silveira (2014), Pilatto & Schumak Melo, Serra (2013), Gargano & Nader (2018) e Couto & Oliveira (2016). Entre estes, os três primeiros apresentam uma visão positiva da jurimetria, exaltando suas potencialidades para a aplicação e para a interpretação do Direito, e os dois últimos, por outro lado, um viés mais negativo, ressaltando problemas na utilização de métodos quantitativos no Direito. Em nenhum dos estudos, porém, adentra-se a uma análise mais aprofundada de sistemas ou de resultados que utilizem técnicas jurimétricas, à exceção do segundo.

Quanto ao texto de Gargano & Nader, que "visa tratar da utilização da Jurimetria nas demandas distribuídas nos juizados especiais cíveis, no qual o objeto da demanda tenha resultado em um pedido de reparação de danos morais" (Gargano & Nader, 2018, .p 18) - assunto este, portanto, que interessa diretamente ao abordado no presente estudo - convém tecer algumas considerações.

Gargano & Nader afirmam que a jurimetria peca por não ter um caráter pessoal, de modo que o uso da estatística em conjunto com o Direito não seria, por este motivo, eficaz "para levantar e compreender os reais motivadores da morosidade excessiva nos processos que tramitam nos Juizados Especiais Cíveis, no que tange à quantificação da valoração do dano moral" (Gargano & Nader, 2018, .p 29). Apesar disso, o trabalho não aponta as situações ou os levantamentos nos quais a alegada falta de pessoalidade da jurimetria teria impactado negativamente ou teria limitado a compreensão de fenômenos jurídicos. As críticas são abstratas, formuladas em tese, o que dificulta sua consideração para fins de avaliar os resultados obtidos com o presente estudo.

Para além das iniciativas que se dedicam a investigar a jurimetria sob um aspecto eminentemente teórico, há aqueles que, como o presente, buscam fazê-lo sob uma perspectiva mais prática. O presente levantamento buscou estudos quantitativos, realizados a partir de dados jurídicos, e foi realizado com o objetivo de aferir, principalmente, a metodologia utilizada nas pesquisas, notadamente quanto às técnicas de aquisição, de enriquecimento, de preparação de dados e de caracterização de dados, os métodos de modelagem de problemas e os recursos utilizados para avaliação dos resultados.

Um dos primeiros estudos efetivamente jurimétricos encontrados foi aquele realizado por Leonardo Netto Parentoni (2012). Para investigar a aplicação prática do instituto da desconsideração da personalidade jurídica, o pesquisador coletou 431 decisões judiciais manualmente nos *sites* de sete tribunais brasileiros, valendo-se de critérios de busca por ele definidos para selecionar apenas itens de interesse, e, a partir da leitura de todos os julgados, extraiu as informações necessárias para realizar análises quantitativas (tais como: classificação dos julgados de acordo com o fundamento utilizado para desconsideração da personalidade jurídica, classificação dos tipos de sociedades que foram alvo da desconsideração, entre outros). Embora certamente inovador na seara jurídica, sobretudo em se tratando de estudo realizado ainda em 2012, os métodos empregados se valeram de poucos recursos computacionais, muito embora as análises quantitativas de fato se valeram de técnicas estatísticas.

Entre os principais estudos de jurimetria no Brasil, destacam-se os realizados pela Associação Brasileira de Jurimetria (ABJ).[1] A instituição vem, desde 2015, atuando isoladamente ou em parceria com outras entidades para a produção de relatórios nos quais analisa dados afetos ao mundo jurídico.

Em geral, os estudos realizados pela associação partem da coleta - que pode ser automatizada (via

[1] https://abj.org.br/

web scrapping)[2] ou não - e do processamento de metadados de processos ou de procedimentos jurídicos para, em seguida, analisá-los à luz de métodos estatísticos. Tais análises podem abranger apenas os referidos metadados[3], ou podem ser enriquecidas com informações qualitativas ou quantitativas obtidas manualmente pelos pesquisadores de alguma forma envolvidos nos estudos (muitas vezes por meio da aplicação de questionários).[4]

Quanto à metodologia utilizada pela associação em trabalhos jurimétricos, realizados em conjunto com outras instituições ou não, em geral são consideradas três etapas - uma de listagem de processos, uma de coleta de dados e outra de análise estatística (Nunes et al., 2022, p. 8) -, com algumas variações a depender das fontes de dados e das análises realizadas. A metodologia geral é explicada em Nunes et al. (2022): a primeira fase consiste na identificação dos processos de interesse, no desenvolvimento dos métodos computacionais para acessar e consolidar informações e na adequação dos dados resultantes ao escopo da pesquisa (Nunes et al., 2022, p. 9-10); a segunda, no preenchimento de formulários (denominados "fichas de classificação") (Nunes et al., 2022, p. 12); e a terceira, na análise propriamente dita (Nunes et al., 2022, p. 13). A segunda fase nem sempre se faz presente nos trabalhos da associação, tendo em vista que, em alguns casos, não há o preenchimento manual de formulários.

Em Nunes (2015), a associação analisou o tempo dos processos relacionados à adoção no Brasil a partir de dados constantes do Cadastro Nacional de Adoção e do Cadastro Nacional de Crianças e Adolescentes Acolhidos, fornecidos pelo Conselho Nacional de Justiça, e de dados extraídos das bases de dados de alguns tribunais brasileiros. O estudo utilizou dados estruturados fornecidos pelas instituições envolvidas, juntamente com dados sobre movimentações processuais coletados por meio de *web crawlers*, sendo que, em parte dos tribunais objeto do estudo, a coleta automatizada não foi possível. Ao final, foram formuladas propostas que buscavam contribuir para uma maior eficiência na tramitação de processos de adoção no Brasil.

Em Nunes & Trenceti (2015), Nunes et al., por meio de *web scrapping*, coletaram informações sobre processos criminais disponíveis no *site* do Tribunal de Justiça de São Paulo (TJSP) com o objetivo de analisar a taxa de reforma de decisões nas câmaras de Direito Criminal do TJSP (Nunes & Trenceti, 2015, p. 3). Neste caso, foram utilizados tanto metadados quanto o texto de decisões proferidas nos processos judiciais analisados.

No caso daquele estudo em particular, os autores extraíram dos textos das referidas decisões os resultados dos processos, organizando-os em quatro grupos: "negaram", "parcialmente", "provido" e "outros". A extração foi feita a partir de *text mining* com o uso de regras lógicas e com a identificação de expressões regulares (Nunes & Trenceti, 2015, p. 5), mas não há maiores detalhes quanto às técnicas empregadas, sendo que o *script* utilizado não se encontra mais disponível para acesso. Em que pese a prática evidencie a existência de um elevado grau de padronização em decisões judiciais, a classificação de textos com base na identificação de expressões regulares pode apresentar dificuldades, na medida em que este tipo de técnica ignora o contexto da informação. Nada garante, por exemplo, que a utilização do termo "parcialmente" ocorra somente em decisões nas quais se deu parcial provimento ao recurso interposto, de modo que a realização de inferências a partir dos resultados precisa considerar as limitações do método empregado. O texto trata do assunto de forma breve, sem apresentar métricas que permitam avaliar a perfomance da tarefa de classificação de resultados realizada, limitando-se a afirmar que "os resultados não são totalmente à prova de erros, mas estamos assumindo que as classificações estão próximas da realidade" (Nunes & Trenceti, 2015, p. 5).

Em relação àquele trabalho, acredita-se que o presente contribui na medida em que busca definir, dentro de seu escopo, um método para que se possa computar o quão confiável é a identificação de expressões regulares em sentenças. Embora as pesquisas tenham escopos e conjuntos de dados absolutamente distintos, a ideia proposta neste trabalho, acredita-se, pode ser replicada em outros que pretendam identificar informações específicas em sentenças judiciais.

Em Nunes & Berger (2020), além de ferramentas de estatística convencional, utilizaram-se três algoritmos diferentes de aprendizado de máquina para análise de dados constantes de processos administrativos sancionadores que tramitaram na Comissão de Valores Mobiliários (CVM): regressão logística, regressão logística com regularização Lasso e florestas aleatórias. Os modelos foram utilizados para que os pesquisadores pudessem identificar as variáveis que influenciavam mais na chance de um processo resultar

[2]Estudos da ABJ em que há coleta de dados via *web scrapping* são: Nunes (2015), Nunes et al. (2014) e Nunes & Trenceti (2015).

[3]Os seguintes trabalhos da ABJ contêm análises de metadados de processos ou de procedimentos jurídicos: Nunes (2016a), Nunes (2015), Nunes et al. (2019) e Nunes et al. (2018).

[4]O recurso à obtenção manual de informações qualitativas ou quantitativas se faz presente em Nunes et al. (2014), Waisberg et al. (2016), Waisberg et al. (2022a), Waisberg et al. (2022b), Nunes et al. (2016), Moisés et al. (2019), Nunes & Berger (2020), Nunes (2020) e Nunes et al. (2022).

em absolvição. A entrada Z era composta por dois conjuntos X e Y, sendo X composto por variáveis com os seguintes metadados sobre os referidos processos: "a existência de termo de compromisso, existência de repercussão pública, existência de defesa, relatores do processo e motivos que fundamentaram a acusação" (Nunes & Berger, 2020, p. 41-42). Y, por sua vez, era um conjunto com apenas uma variável que poderia assumir os valores "Sim", caso o resultado do processo fosse a absolvição, ou "Não", caso não fosse (Nunes & Berger, 2020, p. 41). Z foi dividido em dois subconjuntos de treino e teste na proporção de 80% e 20%, e os melhores modelos (regressão logística com regularização lasso e florestas aleatórias) apresentaram acurácia de 74%(Nunes & Berger, 2020, p. 42). Não foram utilizadas outras métricas de avaliação comuns para algoritmos de classificação.

Outra fonte constante de estudos jurimétricos é o Conselho Nacional de Justiça (CNJ). Além dos estudos feitos em conjunto com a ABJ ou com outras instituições, a entidade publica, todos os anos desde 2003 (Brasil, 2021, p. 9), o relatório Justiça em Números, no qual analisa uma série de metadados relacionados ao Poder Judiciário, de forma mais geral, e aos processos em trâmite nos tribunais brasileiros, em particular. As análises, em geral, são estritamente estatísticas e se valem majoritariamente de uma base de dados elaborada especificamente para o Justiça em Números e de uma base de dados denominada "Base nacional de Dados do Poder Judiciário – DataJud", ambas alimentadas pelos tribunais brasileiros (Brasil, 2021, p. 10). O acesso a tais repositórios não é disponibilizado ao público em geral, muito embora seja possível acessar um *dashboard* que permite a visualização de algumas informações geradas pelo CNJ a partir do DataJud (Conselho Nacional de Justiça, 2022).

Em Unger et al. (2021), os autores utilizam técnicas de mineração de processos para analisar a performance do Tribunal de Justiça de São Paulo a partir de uma base de dados composta de processos relacionados à área do Direito Empresarial. Naquele estudo, a análise recai sobre *logs* de eventos processuais - metadados, portanto - e não adentra no mérito das discussões carreadas nos processos, motivo pelo qual se diferencia substancialmente do presente trabalho.

Por fim, o estudo que mais se relaciona ao presente é aquele realizado por Nunes & Duarte (2021). Em seu levantamento, os autores propuseram um "*modus operandi*" para análise de processos judiciais que versam sobre indenização por dano moral por cadastro indevido em órgãos de proteção ao crédito - assunto, portanto, absolutamente análogo ao aqui tratado (Nunes & Duarte, 2021, p. 470). O objetivo dos autores, ao fazê-lo, é apenas ilustrar um caso de aplicação de jurimetria, que serve como subsídio para uma discussão posterior acerca dos usos da estatística no cotidiano do operador do Direito.

Naquela investigação, os autores coletaram e analisaram manualmente[5] 225 sentenças. Os autores, ainda, definiram um conjunto de 16 variáveis a serem preenchidas com base nas informações disponíveis na sentença e no sistema de buscas do TJMG, as quais foram manualmente coletadas, concomitantemente à coleta. Além disso, também de forma simultânea à coleta e à análise das sentenças, os autores excluíram processos que não se amoldavam aos critérios por eles definidos, o que resultou, ao final, em uma base de dados de 148 sentenças.

Na sequência, Nunes & Duarte buscaram caracterizar sua base de dados, com os objetivos principais de quantificar os pedidos de condenação ao pagamento de danos morais que foram julgados procedentes e os que foram julgados improcedentes e a "distribuição de frequência dos motivos que levaram à improcedência do pedido", bem como definir a moda, a média e a mediana dos valores fixados a título de indenização por danos morais (Nunes & Duarte, 2021, p. 472). Para fins de visualização dos resultados, os autores apresentam gráficos de barras simples, um gráfico de barras empilhadas, um gráfico de pizza e uma tabela (Nunes & Duarte, 2021, p. 472-477).

Em que pese a identidade de objeto de análise, o presente estudo apresenta algumas inovações em relação àquele, sobretudo nas seguintes matérias: coleta automatizada de dados, tratamento de um volume maior de dados com o emprego de recursos computacionais, ampliação das possibilidades de análise estatística, utilização de técnicas de aprendizado de máquina para análise de sentenças e proposição de ferramenta que as utilizem para fins de coleta em trabalhos futuros.

Quanto ao tratamento de um maior volume de dados, este ponto, em específico, foi apontado pelos autores daquele estudo como uma limitação da análise por eles realizada(Nunes & Duarte, 2021, p. 479). Além disso, o sistema do TJMG foi por eles elencado como um dificultador, na medida em que

[5] O texto não deixa claro o método de coleta, limitando-se a afirmar que foram retirados do *site* do TJMG a partir de pesquisa feita em seu sistema de busca de jurisprudência. Em que pese os autores tenham informado os critérios de busca utilizados, a forma pela qual os dados foram efetivamente coletados não é detalhada. Em trecho anterior de seu texto em relação àquele no qual os critérios de busca são descritos, porém, os autores afirmam: "Em assim sendo, se promoveu "manualmente" por cerca de dois meses uma pesquisa jurimétrica acerca do tema proposto de modo a ofertar ao leitor um panorama de como esta abordagem pode ser realizada [...]" (Nunes & Duarte, 2021, p. 470). Além disso, em outro trecho, explica-se que, "a cada sentença analisada", os autores coletaram manualmente informações sobre elas, preenchendo campos por eles pré-definidos (Nunes & Duarte, 2021, p. 471). Assim, assume-se que a coleta se deu manualmente, sem o emprego de *crawlers* ou de outras técnicas que poderiam ser utilizadas.

a coleta de dados, realizada manualmente, gerou grande dispêndio de tempo. (Nunes & Duarte, 2021, p. 480).

Convém, ainda, destacar que o estudo realizado por Nunes & Duarte (2021) não tem como principal objetivo desenvolver ou demonstrar uma metodologia para análise de dados jurídicos, como é o caso desta investigação, mas tão-somente ilustrar um caso de uso para discutir as potencialidades da jurimetria no Direito, destacando sua utilidade para "orientar as partes e seus advogados em negociações" e "no cenário de resolução de disputas *on-line*"(Nunes & Duarte, 2021, p. 479).

Por fim, Nunes & Duarte (2021) ressaltam, ao final de sua análise, que "embora o estudo empírico do Direito não dependa de instrumentos sofisticados, [...] a Jurimetria seria melhor aplicada com auxílio de algoritmos computacionais", o que é justamente o que o presente trabalho se propõe a fazer.

Capítulo 3

Construção da base de sentenças judiciais

Neste capítulo, são descritos os procedimentos utilizados para construção e anotação dos dados utilizados na pesquisa. A tarefa foi executada nas seguintes etapas:

1. **Aquisição das sentenças judiciais**: etapa que compreende uma seleção inicial de processos potencialmente relevantes para o estudo, seguida de *crawling* para obtenção dos textos das sentenças judiciais neles proferidas e da posterior filtragem das sentenças obtidas, a fim de garantir que se amoldavam ao objeto do estudo;

2. **Rotulagem e consolidação da base de dados**: etapa que compreende a identificação manual, em cada uma das sentenças, dos valores fixados a título de indenização por danos morais e de outros valores financeiros nelas mencionados, seguida da consolidação da base de dados.

As tarefas realizadas no âmbito de cada uma destas fases serão detalhadas a seguir.

3.1 Aquisição das sentenças judiciais

Para realização do estudo, gerou-se um *corpus* de sentenças judiciais proferidas em primeiro grau no âmbito do Tribunal de Justiça de Minas Gerais. O *corpus* é composto apenas de decisões que colocam fim à fase de conhecimento nas quais haja a efetiva fixação de danos morais, excluídas decisões proferidas no julgamento de embargos de declaração, decisões proferidas na fase de execução e mesmo sentenças que, embora terminem a fase de conhecimento, deixam de fixar danos morais, seja pela negativa dos pedidos formulados pelo autor, seja pela condenação exclusiva à indenização de danos materiais.

A forma pela qual o referido *corpus* foi construído é detalhada a seguir.

Inicialmente, no âmbito de parceria firmada entre o TJMG e o DCC/UFMG, foi franqueado acesso ao `Radar`, um sistema de recuperação da informação desenvolvido no âmbito daquele tribunal que permite realizar pesquisas em sua base de dados, composta, entre outros, por dados originários do PJe (Diniz et al., 2020, p. 596-597).

Considerando que o presente estudo pretende analisar sentenças que efetivamente fixem danos morais em demandas de negativação indevida, o `Radar` foi utilizado para que fosse possível definir, naquela plataforma, critérios de busca para a seleção de processos que contêm tais decisões. A partir de uma série de testes, foram definidos os seguintes termos de pesquisa: *"danos morais"*, *"negativação indevida"*, *-renúncia*, *-homologo*, *-"sem resolução de mérito"* e *-"487, II"*. Com tais termos foram realizadas duas pesquisas, de modo que, na primeira, foram selecionados apenas sentenças proferidas no âmbito da justiça comum e, na segunda, apenas sentenças proferidas no âmbito dos juizados especiais.

Neste ponto, uma breve consideração é necessária. O `Radar` permite a filtragem dos resultados de acordo com os assuntos discutidos no processo (os quais são escolhidos entre opções pré-definidas pelo Conselho Nacional de Justiça no momento da distribuição da ação, seja pelo advogado, seja pelo responsável pela atermação). Assim, seria possível filtrar apenas processos em que "Inclusão Indevida em Cadastro de Inadimplentes" aparece entre os assuntos discutidos. No entanto, uma análise manual dos processos retornados com e sem a utilização de tal recurso mostrou que existem diversas demandas que, embora versem sobre inscrição indevida em cadastros de inadimplentes, têm seus assuntos rotulados de maneira de incorreta ou simplesmente optam por outros assuntos a elas relacionados, como "Indenização por Dano Moral"e "Contratos Bancários".

Os termos de pesquisa escolhidos tiveram por objetivo excluir da base processos que, embora versassem sobre negativação indevida, não contivessem sentenças que estabelecessem condenações por danos morais. Assim, as expressões -*"sem resolução do mérito"* e -*"487, II"* foram utilizadas para excluir processos encerrados sem julgamento do mérito. A expressão -*"renúncia"* foi empregada para que não fossem selecionados processos nos quais o autor renunciasse à demanda, e -*"homologo"* para excluir processos no âmbito dos quais fosse firmado acordo, de modo que a sentença se limitasse a homologá-lo. Por sua vez, os termos *"danos morais"* e *"negativação indevida"* pretendiam incluir processos que mencionassem estes dois assuntos.

Antes de seguir adiante na exposição, é importante mencionar que, como se pode inferir facilmente, a pesquisa realizada para seleção dos processos tem limitações. A título de exemplo, nada garante que algum processo que efetivamente verse sobre o assunto desejado não tenha sido captado pelos critérios de busca, nem que processos indesejados sejam retornados. Com efeito, na fase de anotação dos dados, foram identificadas uma série de sentenças que não atendem aos objetivos do presente trabalho.

Apesar disso, como este estudo não tem o objetivo de exaurir o universo de sentenças sobre o assunto "Inclusão Indevida em Cadastro de Inadimplentes" e como as sentenças foram, posteriormente, verificadas, uma a uma, de forma manual, não há prejuízo nas falhas decorrentes da busca.[1]

Os critérios de busca definidos foram repassados ao TJMG, que gerou tabelas com informações sobre quatro mil processos retornados a partir da utilização de tais critérios.[2] Os atributos das referidas tabelas são os seguintes:

- *Número CNJ*, que contém o número do processo segundo o padrão do CNJ;
- *Número TJ*, que está vazio para todos os processos da base de dados;
- *Comarca*, que contém a comarca na qual o processo tramita ou tramitou;
- *Natureza*, que assume o valor "CÍVEL" para todos os processos da base;
- *Competência*, que informa o juízo competente para julgar o processo, o qual varia de acordo com a organização judiciária da comarca e com as normas expedidas do TJMG;
- *Classe*, que informa o tipo de procedimento;
- *Assuntos*, indicando os assuntos presentes no processo, informados no momento do cadastro da demanda no PJe;
- *Partes/representantes*, com os nomes das partes envolvidas e seus respectivos advogados (as partes não são vinculadas a um dos polos processuais);
- *Situação*, indicando se o processo se encontrava ativo ou baixado no momento em que a lista foi construída;
- *Meio*, indicando o meio pelo qual o processo tramita (igual a "Eletrônico" para todas as entradas);
- *Sistema*, da qual consta o sistema de tramitação do processo ("PJE" para todas as entradas);
- *Tipo justiça*, que pode ser "Juizado Especial" ou "Justiça Comum";
- *Segredo de justiça*, indicando se o processo está ou não em segredo de justiça, sendo que nenhum dos processos da base está em segredo de justiça;
- *Órgão julgador*, com o nome do órgão responsável pelo julgamento;
- *Data distribuição*, com o carimbo de data e hora nos quais o processo foi distribuído;
- *Tipo distribuição*, indicando se o processo foi distribuído por sorteio ou por dependência;

[1] Hoje, quando um advogado distribui um novo processo judicial no PJe, ele deve classificar o processo de acordo com os assuntos pré-definidos pelo sistema. Nada impede que, ao fazê-lo, o advogado selecione um assunto inadequado. Além disso, há processos que podem se enquadrar em mais de um assunto, de modo que, ainda que todos os processos classificados como relativos a "Inclusão Indevida em Cadastro de Inadimplentes" estivessem corretamente identificados, muito provavelmente existirão outros processos sobre o mesmo assunto rotulados de forma distinta. A criação de ferramentas automatizadas de classificação poderia contribuir para uma seleção mais assertiva de processos com assuntos similares, a partir, por exemplo, do emprego de técnicas de detecção de tópicos e de agrupamento, as quais não são objeto deste trabalho.

[2] No total, foram fornecidas quatro tabelas com quatro mil linhas cada, mas havia processos repetidos, de modo que o número total efetivo de processos, excluídas as repetições, era menor.

- *Magistrado*, com o nome do magistrado responsável pelo processo (alguns processos contêm mais de um magistrado);
- *Julgamentos*, com as datas e as horas nas quais foram proferidas decisões judiciais no processo, incluindo ou não a data da sentença.

Uma vez de posse das tabelas, considerando a impossibilidade de fornecimento, pelo TJMG, das sentenças propriamente ditas, criou-se um *crawler* para automatizar a coleta de tais decisões a partir da página de consulta pública do PJe.[3]

O *crawler* foi construído em `Python` por meio da biblioteca `Selenium` (Muthukadan, 2018), de modo que, para cada linha da planilha, todos os processos foram abertos até a página que revela os andamentos processuais, sendo que, nesta última, foram coletadas as sentenças que aparecessem em primeiro lugar na lista de documentos disponibilizada na primeira página da consulta. Este processo não é perfeito, na medida em que existem documentos rotulados como "sentença" que colocam fim à execução ou que simplesmente alteram a primeira versão da sentença após o julgamento de embargos de declaração.

Cada sentença coletada foi salva em um arquivo de texto (em extensão `.txt`) com uma única linha, cujo título correspondia ao número do processo no qual a sentença foi proferida. Previamente ao salvamento, todos os caracteres de quebra de linha (`\n`) foram retirados, assim como eventuais espaços localizados no começo e no final da `string` resultante. Além disso, os eventuais espaços múltiplos entre caracteres foram reduzidos a apenas um.

Depois, criou-se um pequeno programa que, valendo-se de técnica de identificação de expressões regulares, isolou apenas os processos que continham a `string R$` e que não continham a expressão `NEGO PROVIMENTO`. Ao final, restaram 2.409 processos.

A identificação das referidas expressões regulares, aqui, não gerou qualquer prejuízo à qualidade das análises realizadas posteriormente, na medida em que não teve o objetivo de identificar as sentenças ou as informações objeto de estudo (o que poderia gerar erros), como em (Nunes & Trenceti, 2015), mas tão somente o de remover da base sentenças que poderiam tumultuar a tarefa de anotação manual dos dados, por não conterem valores fixados a título de indenização por danos morais. O único prejuízo gerado por esta abordagem consiste na possibilidade de que sentenças que poderiam ser úteis para a análise tenham sido excluídas, fato este que, no entanto, não teve influência nos resultados aqui apresentados, já que somente foram analisadas as sentenças efetivamente coletadas.

3.2 Rotulagem e consolidação da base de sentenças judiciais

Ao final, após a coleta das sentenças, a exclusão de processos que a princípio não interessavam e a geração de arquivos `.txt`, utilizou-se a biblioteca `label-studio` (Heartex, Inc., 2022) para rotulação dos dados. Foram definidos os seguintes rótulos: `valor_fixado` e `valor_outros`.

O rótulo `valor_outros` identifica todos os valores em reais mencionados no texto da sentença que não correspondem ao montante fixado pelo juízo a título de indenização por danos morais. Exemplos de valores que receberam este rótulo podem ser verificados na Tabela 3.1.

Por sua vez, o rótulo `valor_indenizacao` identifica todos os valores em reais mencionados no texto da sentença que efetivamente correspondem ao montante fixado pelo juízo a título de indenização por danos morais. Exemplos de valores que receberam este rótulo podem ser verificados na Tabela 3.2.

Todos os valores constantes das sentenças foram rotulados em uma das duas categorias. Além disso, ainda que um mesmo valor fosse mencionado mais de uma vez, ele foi rotulado em todas as suas ocorrências, como no segundo exemplo da Tabela 3.2.

Além disso, concomitantemente ao processo de rotulação, foram excluídas sentenças que não condenavam a parte ré a reparar danos morais sofridos pela parte autora ou que fixassem valores escritos apenas em formato não numérico ou em formato numérico distinto do usualmente utilizado para representar quantias em reais. Assim, por exemplo, sentenças que fixavam condenações em "7 mil reais" ou

[3] A página de consulta pública do PJe permite que qualquer pessoa consulte algumas informações sobre processos judicias que não tramitam em segredo de justiça, como seu andamento e alguns dos documentos que o compõem, incluindo as decisões judiciais que nele foram proferidas. No caso do TJMG, pode ser acessada a partir da seguinte *URL*: `https://pje-consulta-publica.tjmg.jus.br/`.

Tabela 3.1. Valores rotulados como "outros".

Trecho	*Valores rotulados*
O documento de ID206335215 demonstra que o saldo devedor referente ao contrato nº2296329 era de `R$442,70` e foi quitado em 10/10/2019. E ainda, o documento de ID 206335215 demonstra que o saldo devedor referente ao contrato nº2136808 era de `358,44` e foi quitado em 10/10/2019.	442,70 e 358,44
O negócio jurídico foi firmado no primeiro semestre de 2014 e sofreu uma renegociação no segundo semestre do mesmo ano, nesse sentido o valor total do contrato era de `R$ 59.321,90`.	59.321,90

Tabela 3.2. Valores rotulados como "indenização".

Trecho	*Valores rotulados*
Em face do exposto, julgo parcialmente procedente o pedido do Requerente. 3. DISPOSITIVO Em face do exposto, julgo procedente o pedido do Requerente, condenando a Requerida ao pagamento de `R$ 10.000,00` (dez mil reais), a título de danos morais.	10.000,00
No presente caso, não há provas de que houve circunstância extraordinária à própria negativação, pelo que tenho por razoável a fixação do valor em `R$ 6.000,00` (seis mil reais) a título de reparação dos danos morais. Pelo exposto, e por tudo o mais que dos autos consta, julgo PROCEDENTES os pedidos contidos na inicial, com força no artigo 487, I, do Código de Processo Civil, para declarar inexistente o débito entre as partes e para condenar a requerida a pagar à requerente o valor de `R$ 6.000,00` (seis mil reais) a título de danos morais, devidamente corrigido, desde o arbitramento (súmula 362 do egrégio STJ), pelos índices da Corregedoria Geral de Justiça deste Estado e com juros de mora desde o evento danoso (súmula 54 do egrégio STJ), ou seja, a inscrição indevida.	6.000,00 e 6.000,00

em "sete mil reais" foram desconsideradas, sendo aceitos, por outro lado, formatos como "`R$ 7.000,00` (sete mil reais)". Em casos como este último, apenas os números e símbolos existentes entre eles foram considerados.

Foram excluídas da base de dados utilizada para anotação, ainda, sentenças com mais de um valor de indenização e sentenças proferidas no âmbito do julgamento de embargos de declaração, no âmbito da fase de cumprimento de sentença ou para fins de homologar acordo firmado entre as partes. A análise que culminou na exclusão de tais sentenças foi realizada manualmente, caso a caso.

Vale ressaltar que `strings` relativas a valores precedidas por mais de uma `string` `R$` (com ou sem espaços) ou que possuíam menos de dois zeros após a vírgula também foram manualmente rotuladas, assim como aquelas que não eram precedidas de `R$`, que eram precedidas apenas de `R` ou apenas de `$` ou que, embora sem zeros após a vírgula ou sem estarem precedidas de quaisquer dos caracteres `R` ou `$`, pudessem ser categorizadas, em virtude do contexto, de acordo com algum dos dois rótulos.

Ao final, foram anotadas 1.422 sentenças com todos os seus valores devidamente rotulados em uma das duas categorias mencionadas. Vale ressaltar que todas as sentenças têm, pelo menos, um valor com o rótulo `valor_indenizacao`, já que a existência de condenação era um dos pressupostos da montagem da base de dados.

As sentenças anotadas foram então exportadas em formato `.json`, e, como estavam indexadas pelos números dos processos em que foram proferidas tanto nos nomes dos arquivos quanto nas planilhas fornecidas pelo TJMG, os dados do referido arquivo foram cruzados com os metadados daquelas planilhas, gerando a base de dados utilizada para as fases seguintes do trabalho, composta pelas seguintes colunas:

- *processoNum*, com os números dos processos segundo o padrão do CNJ;
- *sentencaTexto*, com os textos integrais das sentenças;
- *sentencaAnotacoes*, que contém as anotações feitas para cada sentença;
- *processoCompetencia*, cujo valor pode ser "Juizado Especial" ou "Justiça Comum";
- *processoComarca*, com as comarcas nas quais os processos tramitaram;
- *processoJuizo*, com o nome do órgão responsável pelo julgamento;
- *sentencaMagistrado*, com o nome do magistrado que proferiu a sentença;
- *processoPartes*, com os nomes das partes e advogados envolvidos no processo;
- *processoDistribuicao*, com as datas de distribuição das ações;
- *sentencaData*, com as datas nas quais as sentenças judiciais foram proferidas.

Alguns campos que estavam em branco nas tabelas inicialmente fornecidas pelo TJMG, especificamente nas colunas que identificavam o magistrado responsável pelo processo e a data da sentença, tiveram de ser preenchidos manualmente. As informações foram retiradas do próprio texto da sentença, no caso da identificação do magistrado, ou do sistema de consulta pública ao acompanhamento processual do PJe. Além disso, algumas células da tabela inicialmente fornecida pelo TJMG continham mais de um valor no campo "Magistrado". Por este motivo, utilizou-se uma heurística pela qual, para cada um dos nomes, verificou-se se constava da sentença, partindo de trás para frente. Em caso afirmativo, o primeiro nome encontrado era adicionado à base de dados final, com a exclusão dos demais. Nos casos em que nenhum foi encontrado, o que aconteceu raras vezes, o preenchimento foi feito de forma manual.

Capítulo 4

Caracterização da base de sentenças judiciais

Este Capítulo apresenta as técnicas utilizadas para caracterização da base de sentenças judiciais. Trata-se de mera caracterização, sendo que uma discussão mais aprofundada em relação aos resultados somente será feita no Capítulo 6.

4.1 Informações gerais sobre a base de dados

A base de dados possui as seguintes características gerais, aplicáveis tanto a processos originários da justiça comum quanto a processos oriundos dos juizados especiais cíveis:

Tabela 4.1. Características da base de dados.

Característica	*Valor*
# Total de sentenças	1.422
# Processos nos juizados especiais	717
# Processos na justiça comum	705
# Diferentes valores de indenização	36
# Diferentes comarcas	158
# Diferentes magistrados	514
Intervalo de datas de distribuição	24/08/12 a 08/02/22
Intervalo de datas de juntada de sentenças aos autos	04/08/16 a 05/05/22

A indicação dos intervalos tem somente o objetivo de auxiliar na compreensão da base de dados. Assim, não é apenas possível como altamente provável que, no mesmo intervalo de tempo aqui exposto, tenham sido proferidas inúmeras sentenças que, por não constarem da planilha enviada pelo TJMG ou em virtude de restrições ou limitações das técnicas empregadas na fase de coleta de dados descrita anteriormente, não constam da base de dados analisada.

Além disso, é importante ressaltar que, embora contenha cerca de dez vezes mais decisões que as analisadas em Nunes & Duarte (2021), a base de dados apresenta um recorte muito restrito da realidade do TJMG. Não se pode dizer que o número de processos representa a totalidade dos que tramitam naquele tribunal. Além disso, o número de comarcas representa cerca de metade de todas as que existem no estado de Minas Gerais. Assim, é preciso ter cautela na análise dos resultados expostos a seguir: embora seja possível formular conjecturas e inferir tendências a partir dos dados analisados, generalizações devem ser evitadas.

4.2 Análise estatística

4.2.1 Valores de indenização

Quanto aos valores fixados a título de indenização por danos morais nas sentenças constantes da base, foram computadas a *média*, a *mediana* e a *moda* para todo o conjunto de dados (Magalhães & Lima, 2015, p.106), com os resultados evidenciados na Tabela 4.2.

Tabela 4.2. Medidas de posição para os valores de indenização.

Medida	*Valor*
Média	6.297, 47
Mediana	6.000, 00
Moda	6.000, 00

O desvio padrão (Magalhães & Lima, 2015, p.116-117) dos valores das indenizações constantes da base é de aproximadamente `R$ 3328,45`.

Quando separados de acordo com a competência (justiça comum ou juizado especial), os valores fixados pelos magistrados apresentam as distribuições evidenciadas nas Figuras 4.1 e 4.2.

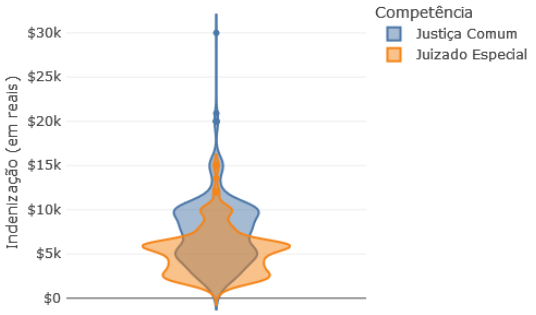


Figura 4.1. Gráfico de violino que evidencia as distribuições dos valores fixados pelos magistrados de acordo com a competência.

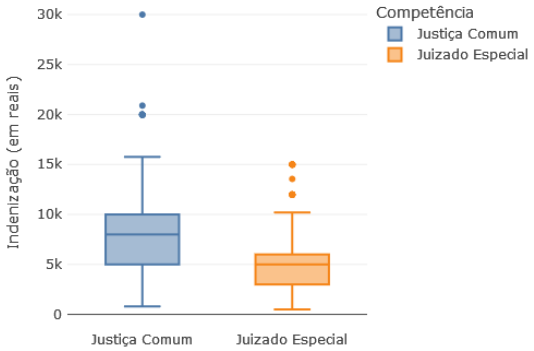


Figura 4.2. Diagrama de caixa que evidencia as distribuições dos valores fixados pelos magistrados de acordo com a competência.

Como se vê na Figura 4.1, a curva que representa a distribuição dos valores fixados a título de indenização por danos morais no juizados é mais achatada que a curva relativa à justiça comum, sendo que, na segunda, os valores tendem a ser maiores. A mesma tendência se verifica na Figura 4.2, que deixa mais evidentes as diferenças nos primeiros, segundos e terceiros quartis para cada uma das competências. A medianas dos valores de indenização fixados na justiça comum e dos fixados no âmbito dos juizados especiais são, respectivamente, `R$ 5.000,00` e `R$ 8.000,00`. O coeficiente de assimetria de Bowley (Magalhães & Lima, 2015, p.25) é de $-0,2$, no primeiro, é de aproximadamente $0,33$, no segundo, o que indica que, em ambos, valores menores aparecem com maior frequência que valores maiores. Quando comparados os dois, no entanto, tem-se que os valores praticados no primeiro são maiores que os verificados no segundo, o que é facilmente perceptível pelo gráfico.

Naquela figura, nota-se, ainda, que a variabilidade de valores de indenização fixados na justiça comum é maior que a verificada nos juizados especiais, sendo que o número de elementos em cada um

deles é bastante similar (o grupo de processos que tramitaram nos juizados especiais tem apenas 12 sentenças a mais que o grupo de processos que tramitaram na justiça comum).

A tendência de valores maiores na justiça comum se mantém ao longo dos anos, conforme evidenciado nas Figuras 4.3 e 4.4. Para os valores fixados no âmbito dos juizados especiais, a curva de tendência evidenciada na Figura 4.4 é praticamente paralela ao eixo `x`, denotando uma manutenção dos valores praticados ao longo do tempo para os processos constantes da base. Já quando considerados os processos que tramitaram na justiça comum, nota-se uma tendência de queda nas datas mais recentes.

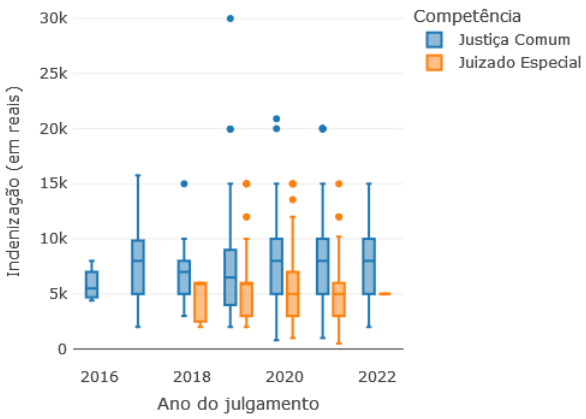


Figura 4.3. Diagrama de caixa dos valores de indenização por competência por ano de juntada da sentença aos autos do processo.

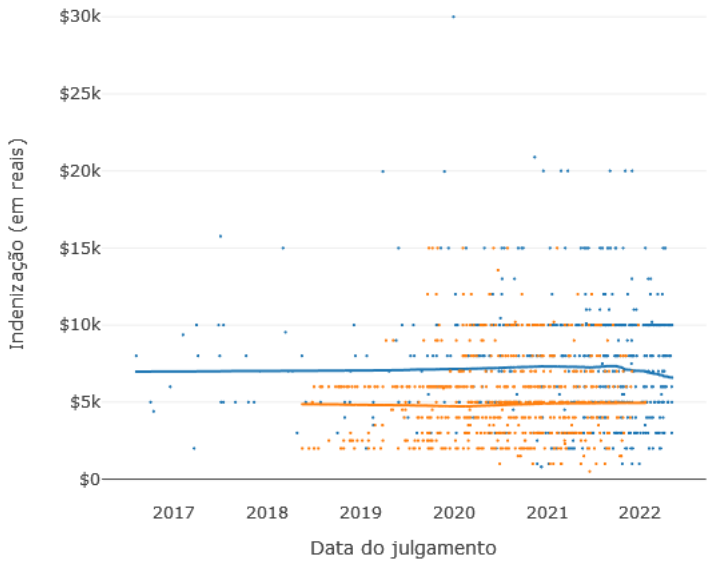


Figura 4.4. Gráfico de dispersão com os valores de indenização fixados ao longo do tempo por competência, com curva de tendência calculada pelo método Lowess.

Se consideradas as três comarcas mais representadas na base de dados (Betim, com 247 sentenças, Belo Horizonte, com 239 sentenças, e Contagem, com 96 sentenças) o padrão de valores maiores na justiça

comum em relação ao praticado nos juizados especiais se mantém, conforme evidenciado na Figura 4.5. Nesta caso, vale ressaltar que existe uma assimetria entre os processos constantes de cada um dos grupos induzidos pelas referidas comarcas: entre os processos que tramitaram na comarca de Contagem, 63 o fizeram sob o rito dos juizados especiais, contra 33 sob o rito da justiça comum; em Betim, foram 214 no juizado contra 33 na justiça comum; e, em Belo Horizonte, foram 70 sob o rito dos juizados e 169 sob o da justiça comum.

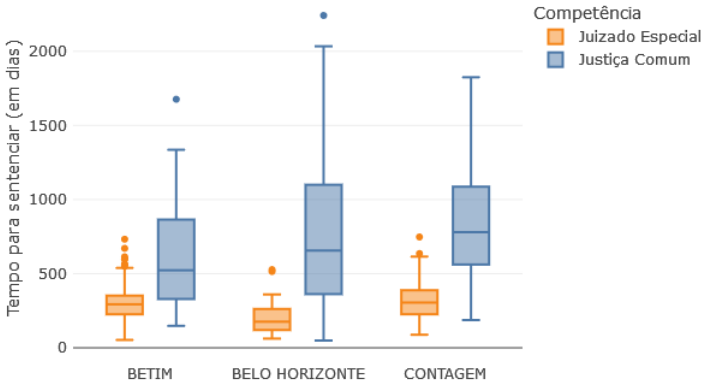


Figura 4.5. Diagrama de caixa dos valores de indenização praticados nas três comarcas com maior número de sentenças, por competência.

Quando analisados os valores fixados por juízes, é possível identificar alguns padrões. Na Figura 4.6 estão representados os valores fixados a título de indenização por danos morais por magistrado que seja responsável por, pelo menos, 25 sentenças na base de dados. Cada coluna em x representa um magistrado distinto, cujos nomes foram substituídos por *strings* de 3 letras maiúsculas ou números aleatórios, posto que os nomes reais são irrelevantes para os fins deste estudo.

Os valores fixados pelos três primeiros magistrados representados na Figura 4.6, todos pertencentes à comarca de Betim, têm suas medianas fixadas em `R$ 6.000,00`, sendo que o magistrado `J7O` possui 87 sentenças na base, o `CZV` possui 66, e o `E4C` possui 58. Os *box plots* relativos aos três magistrados, por sua vez, são absolutamente similares, indicando um alto grau de padronização nos valores por eles fixados.

Há apenas um magistrado com mais de `25` sentenças na base que atuou em duas comarcas distintas (com apenas um processo na competência "Justiça Comum"). Trata-se do magistrado `G1O` representado na Figura 4.6, cujos valores fixados a título de indenização por danos morais são sempre de `R$ 4.000,00`, independentemente do processo.

Há, no entanto, um magistrado cujos valores variam muito, qual seja, o magistrado `62D`. Analisando uma amostra aleatória de quatro decisões por ele proferidas, não foi possível identificar os motivos que levaram a tal variação. Na referida amostra, duas decisões tiveram condenações no montante de `R$ 15.000,00`, uma no valor de `R$ 8.000,00` e a última no de `R$ 5.000,00`, sendo que apenas em uma houve um critério explicitamente considerado para calcular o montante da indenização, fixando-o no maior patamar identificado para este magistrado: o fato de o autor ser pessoa idosa.

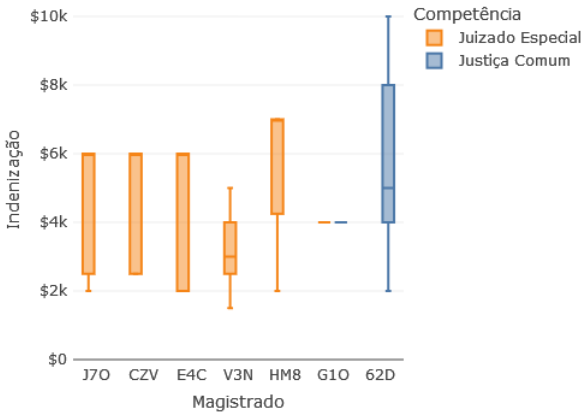


Figura 4.6. *Box plot* dos valores de indenização fixados por magistrado por competência.

4.2.2 Valores de indenização praticados para as partes mais frequentes

Para fins de se identificar as partes mais frequentes na base de dados,[1] buscou-se agrupar partes integrantes de um mesmo grupo econômico. Por exemplo, considerou-se que as partes "TIM CELULAR SA", "TIM SA", "TIM" todas fazem referência a empresas pertencem a grupo econômico que controla e utiliza a marca "TIM", que identifica serviços e produtos de telecomunicações. Tal agrupamento é relevante porque, no Brasil, é muito comum que grupos econômicos sejam constituídos por diversas pessoas jurídicas distintas, as quais utilizam ou fazem referência às mesmas marcas de produtos ou serviços, de modo que, do ponto de vista do consumidor, o fato de se estar tratando com a pessoa jurídica "TIM SA" ou com a pessoa jurídica "TIM CELULAR SA" pouco importa, já que quem aparece para ele, de forma mais intuitiva, é sempre a "TIM".

O trabalho de agrupar as partes por grupo econômico foi feito de forma semi-automatizada, estabelecendo-se como linha de corte as partes que aparecem em mais de 40 processos constantes da base de dados.[2] Para cada uma delas, identificou-se uma *substring* que identifica a marca a elas associada. Depois, seguiu-se à identificação de todas as demais partes cujos nomes continham aquela mesma *substring*. No caso da marca "VIVO", por exemplo, todas as empresas com a *substring* "VIVO" foram agrupadas sob o nome "VIVO". No caso da marca "OI", foram agrupadas também partes cujos nomes, embora não contivessem a *substring* "OI", continham as *substrings* "TELEMAR" e "NORTE LESTE", que sabidamente pertencem ao mesmo grupo econômico. O mesmo se fez com as *substrings* "TELEFONICA", "NET", "ITAUCARD" e "BRADESCARD", as quais identificam partes integrantes, respectivamente, dos grupos "VIVO" , "CLARO", "ITAU" e "BRADESCO". É possível, vale ressaltar, que o número real de sentenças por parte ou por grupo seja maior que o apresentado, já que podem existir outras empresas que poderiam ser agrupadas cujos nomes não foram identificados.

As partes mais frequentes obtidas deste processo, entre as 1.761 possíveis, estão listadas na Tabela 4.3. De todo o universo de 1.422 processos, três dos quatro maiores grupos econômicos de telecomunicações do Brasil figuram como partes em cerca de `34,95%`, enquanto os quatro grupos de empresas do setor

[1] Em se tratando de jurimetria, uma tarefa interessante é compreender se há diferenças na atividades jurisdicional em relação a diferentes réus. Por um lado, a imparcialidade exigida do Poder Judiciário dita que o tratamento entre diferentes partes deve ser igualitário, de modo que, para uma situação similar, o fato de o réu ser uma ou outra pessoa, física ou jurídica, não pode determinar tratamentos distintos. Por outro, aferir quem é mais demandado pode (ou deveria) impactar na fixação de indenizações, tendo em vista que, conforme anteriormente exposto, um dos critérios para definição do valor é a necessidade de coibir a prática de novos ilícitos. Além disso, compreender as partes ou setores mais demandados pode contribuir para a formulação de políticas públicas ou de regulamentações setoriais.

[2] O número 40 foi definido como linha de corte de maneira arbitrária, considerando a dificuldade de se agrupar as partes de maneira semi-automatizada, com boa parte da tarefa tendo de ser executada manualmente.

bancário listados participam de aproximadamente 21,02% das ações.

Embora a base de dados fornecida pelo TJMG não contenha esta informação, presume-se que as partes elencadas na Tabela 4.3 figuram no polo passivo dos processos a que se referem. Tal presunção decorre do relatório feito por Associação Brasileira de Jurimetria (2017), no qual se constatou a existência, nas bases de dados nele analisadas, de uma concentração de litigantes em demandas consumeristas dos setores bancário e de telecomunicações (Associação Brasileira de Jurimetria, 2017, p.122).

Tabela 4.3. Lista das partes mais frequentes na base de dados. De todas, apenas a última não foi objeto de agrupamento.

Parte	*# Processos*
VIVO	323
BRADESCO	142
OI	132
ITAU	56
SANTANDER	56
BANCO DO BRASIL	45
TIM	42

Agrupadas por setor de mercado, tem-se que os valores de indenização praticados quando as maiores litigantes do setor de telecomunicações estão presentes no processo são, em geral, menores que aqueles fixados se presentes empresas do setor bancário. É o que se depreende da Figura 4.7.

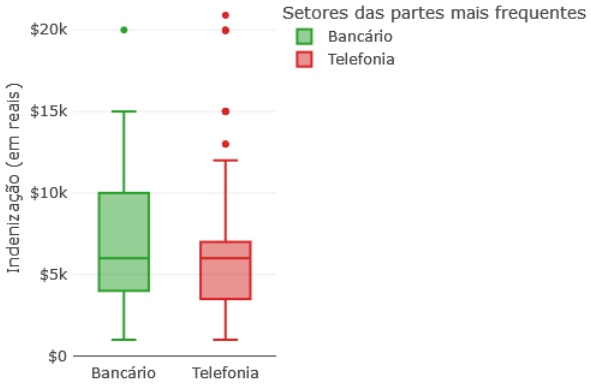


Figura 4.7. Diagrama de caixa representando os valores de indenização fixados para cada um dos setores a que pertencem as partes mais frequentes na base de dados.

Como se nota na Figura 4.7, embora as medianas sejam as mesmas para os dois grupos e os primeiros percentis sejam praticamente idênticos, o fato de o terceiro percentil estar muito mais próximo da mediana no grupo "Telefonia" denota uma maior concentração de valores em patamares mais baixos neste que no outro grupo.

Se analisados os valores fixados para cada uma das partes mais frequentes por competência, em geral mantém-se o padrão verificado anteriormente de valores mais baixos nos juizados especiais que na justiça comum, conforme evidenciado na Figura 4.8.

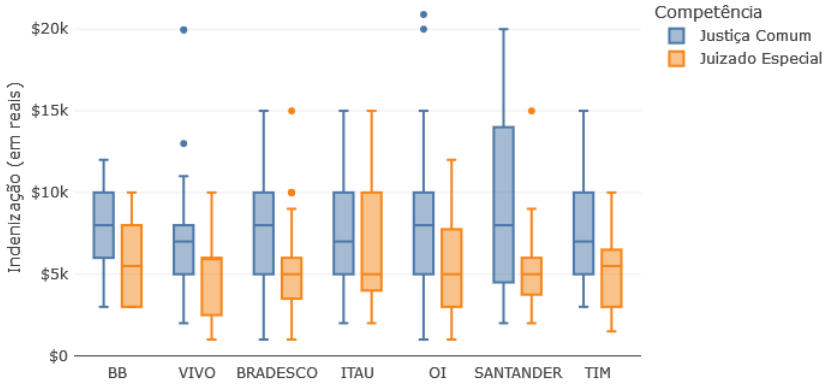


Figura 4.8. Diagrama de caixa representando os valores de indenização fixados por competência para cada uma das partes mais frequentes na base de dados.

4.2.3 Intervalo de tempo entre a distribuição e a juntada da sentença ao processo

Na sequência, foram computadas a *média* e a *mediana* (Magalhães & Lima, 2015, p.106) de todos os intervalos de tempo transcorridos entre as datas de distribuição de cada um dos processos constantes da base de dados e as datas de juntada de suas respectivas sentenças, com os resultados evidenciados na Tabela 4.4. O desvio padrão verificado foi de aproximadamente $395,46$ dias.

Tabela 4.4. Medidas de posição para os tempos transcorridos entre a distribuição e a juntada da sentença aos autos, considerando todos os processos da base de dados.

Medida	*Tempo (em dias)*
Média	$461,33$
Mediana	325

Quando analisados os períodos de tempo transcorridos entre a distribuição das ações e a juntada das sentenças aos autos do processo, por competência, tem-se os gráficos evidenciados nas Figuras 4.9 e 4.10.[3]

Como se nota (e como é de se esperar, dadas as normas que disciplinam o procedimento dos juizados especiais cíveis),[4] quando considerada toda a base de dados, os períodos são maiores na justiça comum que nos juizados especiais. Com efeito, na Figura 4.9, o gráfico que representa os processos que tramitaram nos juizados especiais é muito mais achatado que aquele que representa os processos de competência da justiça comum. Além disso, as medianas do primeiro e do segundo grupo diferem por 314 dias (o que corresponde a mais de dez meses de diferença), e as médias dos valores fixados em cada caso

[3] O tempo transcorrido entre a distribuição das ações e a juntada das sentenças aos autos do processo é denominado, nos gráficos aqui apresentados, como "tempo para sentenciar". Não se considerou a data constante do texto da sentença para os fins deste estudo em virtude, principalmente, de dois fatores. Em primeiro lugar, é possível que o texto da sentença não contenha a data em que foi proferida. Em segundo lugar, é mais fácil extrair as datas dos andamentos processuais (que dão conta apenas da data em que a sentença foi juntada aos autos, não necessariamente daquela em que foi redigida pelo juiz) que dos textos das sentenças em si.

[4] Como anteriormente explicitado nesta dissertação, o procedimento dos juizados especiais cíveis é composto por regras elaboradas para deixá-lo mais simples e rápido que o procedimento comum, aplicável às demandas de negativação indevida que tramitam na justiça comum. Apesar disso, não há um prazo definido para a duração de processos judiciais, então é possível que processos que tramitam na justiça comum sejam mais rápidos que aqueles que tramitem nos juizados. Esta possibilidade, aliás, é facilmente verificada na Figura 4.13, na qual existem pontos azuis (que indicam processos que tramitaram na justiça comum) situados abaixo de pontos laranjas (que representam processos que tramitaram segundo o rito dos juizados especiais).

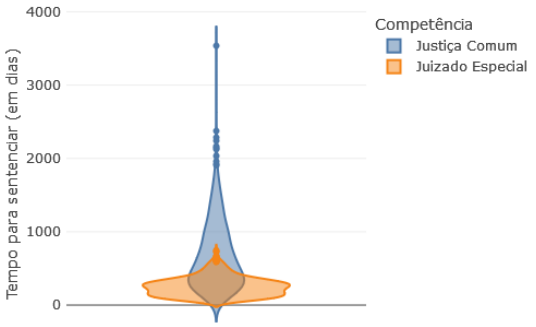


Figura 4.9. Gráfico de violino que evidencia as distribuições dos tempos transcorridos entre a distribuição da demanda e a juntada da sentença aos autos de acordo com a competência.

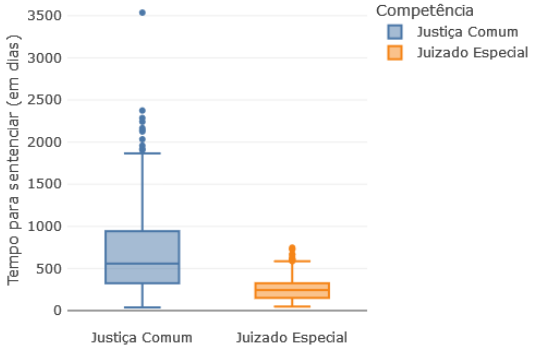


Figura 4.10. Diagrama de caixa que evidencia as distribuições dos tempos transcorridos entre a distribuição da demanda e a juntada da sentença aos autos de acordo com a competência.

(de 673, 55 dias para a justiça comum e de 252, 65 para os juizados) têm uma diferença de 420, 9 dias - mais de um ano, portanto. Em geral, dado um mesmo número de dias transcorridos desde a propositura da demanda, a probabilidade de a sentença já ter sido juntada aos autos do processo é maior no juizado especial que na justiça comum, conforme evidenciado na Figura 4.11.

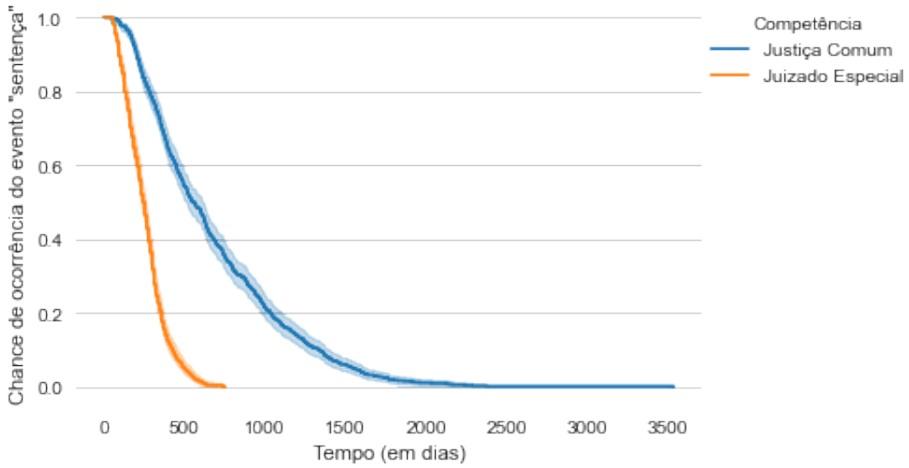


Figura 4.11. Curva de sobrevivência considerando o tempo transcorrido desde a propositura da demanda e o evento "juntada da sentença aos autos".

Isto significa que, se o objetivo de uma determinada pessoa cujo nome seja negativado indevidamente for minimizar o tempo necessário para solucionar sua demanda perante o Poder Judiciário, este estudo dá pistas de que o melhor caminho é, provavelmente, recorrer aos juizados especiais - embora os valores de indenização nele praticados sejam geralmente menores.

A mesma relação (relativa à existência de períodos de duração maiores na justiça comum do que nos juizados especiais) se verifica nas três comarcas com maior número de processos na base, conforme evidenciado na Figura 4.12, muito embora a distância entre os períodos de duração para cada competência varie de uma comarca para a outra, sendo a maior discrepância verificada na comarca de Belo Horizonte (que também é a que possui os maiores períodos), o que talvez se justifique pelo tamanho da comarca e pelo grande número de demandas que nela tramitam.

A verificação de períodos menores nos juizados especiais persiste ao longo de tempo, conforme evidenciado pela Figura 4.13. Nela, vê-se que a linha de tendência dos processos julgados nos juizados

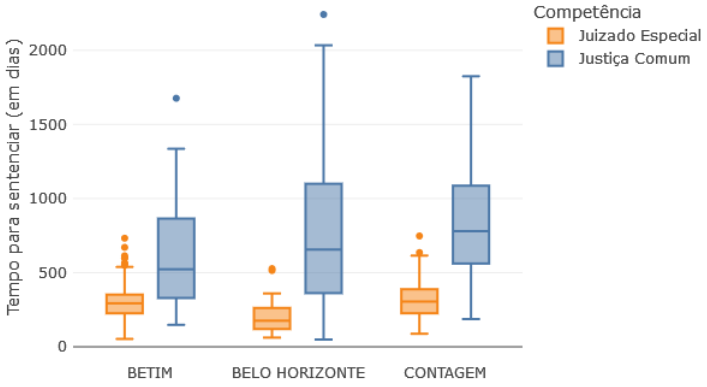


Figura 4.12. *Box plot* dos tempos transcorridos entre a distribuição da ação e a juntada da sentença aos autos nas três comarcas com mais processos na base de dados.

especiais é praticamente paralela ao eixo `x`, porquanto a linha de tendência para as durações dos processos julgados na justiça comum apresenta uma elevação e uma queda nas datas mais recentes. O motivo para isso é desconhecido, mas, de todo modo, a variação parece pouco relevante diante do todo, de modo que, também para a justiça comum, a tendência parece ser de continuidade.

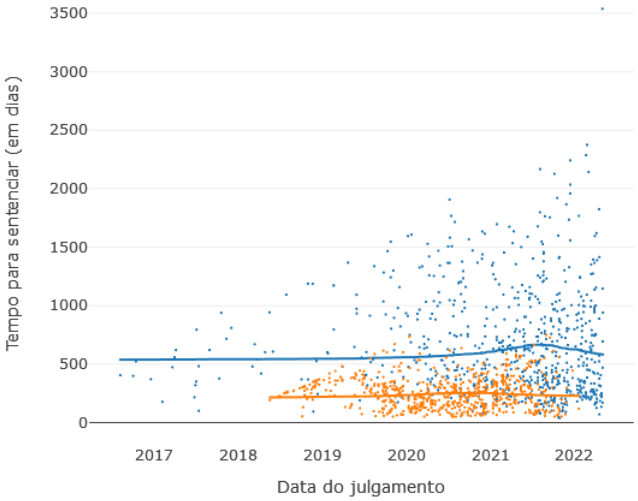


Figura 4.13. Gráfico de dispersão com os tempos transcorridos entre a distribuição dos processos e a juntada das sentenças aos autos por competência, com curva de tendência calculada pelo método Lowess.

Em seguida, passou-se à caracterização dos períodos decorridos entre a distribuição e a juntada da sentença aos autos por magistrado com mais de 25 sentenças na base de dados (Figura 4.14). Como se nota, a tendência de se verificarem tempos maiores na justiça comum que nos juizados, em geral, se mantém. Vale frisar, neste ponto, que apenas a partir das sentenças não se pode extrair grandes conclusões acerca dos motivos pelos quais um magistrado demora mais que o outro, sobretudo dentro de um mesmo grupo (Juizado Especial ou Justiça Comum), análise esta que demandaria outras informações que não foram consideradas nesta pesquisa, tais como: grau de complexidade dos diferentes processos, atuação das partes, atuação das secretarias dos juízos, nível de congestionamento dos diferentes juízos,

entre outros aspectos possivelmente relevantes.

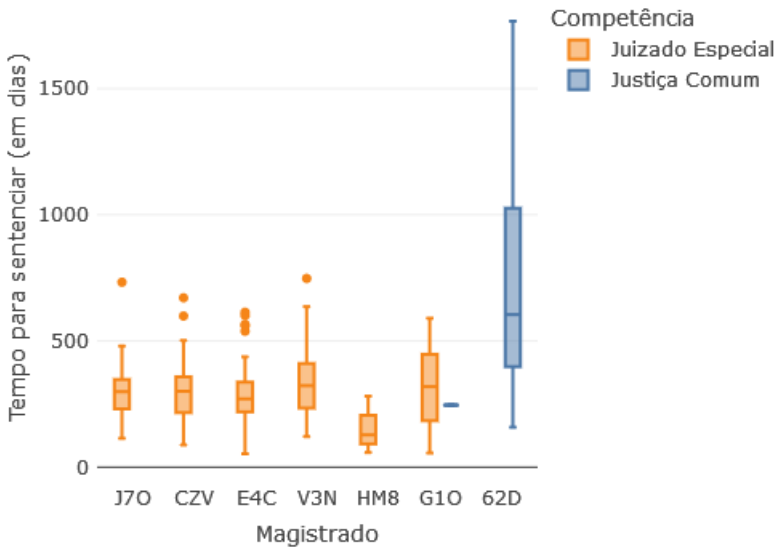


Figura 4.14. *Box plot* dos tempos transcorridos entre a distribuição da ação e a juntada da sentença aos autos por magistrado com pelo menos dez sentenças.

De todo modo, a tendência de se verificarem períodos maiores e indenizações maiores na justiça comum - e, consequentemente, períodos menores e indenizações menores nos juizados especiais - revela um *tradeoff* entre a maximização do resultado financeiro decorrente do processo e a minimização do tempo para sua obtenção.

4.3 Da caracterização à modelagem

Até aqui, com a análise estatística dos dados, é possível confirmar, *para as sentenças constantes da base analisada*, a hipótese segundo a qual os valores praticados na justiça comum são maiores que aqueles fixados nos juizados especiais. Tal relação se manteve em todos os recortes feitos para a avaliação dos dados, independentemente da data da prolação da sentença, da comarca na qual foi proferida e do réu que figurava no polo passivo da demanda.

A avaliação das demais hipóteses, no entanto, depende da modelagem dos dados, descrita no capítulo seguinte, a qual se tornou possível por uma característica importante das sentenças judiciais aqui analisadas: sua homogeneidade no que diz respeito ao contexto em que se inserem. Com efeito, as variações, ao longo do tempo, nos valores fixados e nos tempos de duração dos processos judiciais são, em geral, pequenas, tanto na justiça comum quanto nos juizados especiais, sendo que as diferenças verificadas entre ambas as competências, caracterizadas pelos maiores valores e prazos de duração na justiça comum que nos juizados especiais, restaram presentes independentemente de variáveis como o tempo, a comarca e os réus envolvidos nos processos.

Capítulo 5

Uso de técnicas de aprendizado de máquina para modelagem de dados

Uma vez caracterizadas as sentenças judiciais sob uma perspectiva estatística, passou-se à utilização de algoritmos de aprendizado de máquina com finalidades distintas.

Em primeiro lugar, a fim de investigar a similaridade entre as sentenças e a possibilidade de agrupá-las de modo que sentenças com valores de indenização similares ficassem em um mesmo grupo, foram utilizados três algoritmos de agrupamento distintos: `K-Means`, `Agglomerative Clustering` e `Gaussian Mixture`. O experimento é descrito na Seção 5.3. Nesta fase, buscou-se aliar análises quantitativas e qualitativas, de modo a favorecer a compreensão dos resultados.

Na sequência, buscou-se aferir a possibilidade de, dada uma sentença, prever o valor fixado a título de indenização. Para tanto, utilizaram-se dois algoritmos de regressão, `SVR` e `SGD`, além de uma regressão linear para fins de traçar uma linha de base. O experimento é descrito na Seção 5.4.

Para alimentar os algoritmos de agrupamento e de regressão, as sentenças judiciais foram previamente submetidas a um processo de tokenização[1] e de vetorização[2], conforme exposto a seguir.

Mais uma vez, frisa-se que esta seção tem o principal objetivo de descrever a metodologia utilizada e de expor os resultados, sendo que a discussão posterior dos achados é realizada no Capítulo 6.

5.1 Tokenização e vetorização

Inicialmente, as sentenças integrantes da base de dados foram submetidas a um processo de *tokenização*. Em primeiro lugar, foi definida uma lista de *stopwords* a partir da composição entre a lista padrão da biblioteca `NLTK` (Bird et al., 2009) e uma lista definida com base na análise qualitativa dos textos. Na sequência, os textos foram pré-processados para conversão de letras em caixa alta para caixa baixa, remoção de *URLs*, de pontuação e de outros símbolos (como `$`). Além disso, foram removidos alguns trechos específicos que em nada contribuem para a compreensão dos textos (tais como "id documento imprimir gerar pdf", "assinado eletronicamente por", "poder judiciario" e "estado minas gerais")[3] e os valores fixados a título de indenização por danos morais (tanto os números propriamente ditos quanto seu equivalente por extenso, já que a repetição do valor em moeda, por extenso, é comum em textos jurídicos). A remoção dos valores foi feita de maneira automática a partir do reconhecimento de expressões regulares, notadamente, no caso de valores por extenso, do reconhecimento de expressões entre parêntesis.

Cada uma das sentenças passou, então, por um processo de lematização (Ozturkmenoglu & Alpkocak, 2012), realizado com o uso da biblioteca SpaCy. Assim, cada palavra foi reduzida a seu lema, de modo que cada sentença passou a ser representada como uma sequência de lemas sem pontuação. Além disso, foram excluídos *tokens* que aparecem em mais de 98% e em menos de 2% das sentenças. Como resultado, restou um vocabulário de tamanho $|V| = 2.885$.

[1] Todo *pipeline* de processamento de linguagem natural começa com um processo denominado *tokenização*(Lane et al., 2019, p.33), que consiste na transformação das palavras integrantes das sentenças que compõem a base de dados em *tokens*.

[2] *Vetorização* é o processo que converte as sentenças já "tokenizadas" em vetores numéricos.

[3] A lista de expressões removidas é a seguinte: "poder judiciario estado minas gerais", "tribunal justica minas gerais", "assinado eletronicamente por", "id documento imprimir gerar", "assinado eletronicamente", "poder judiciario", "comarca", "recursoprocesso", "cep", "que", "id", "sa", "se", "id", "pois", "modo", "alem", "sob", "c/c", "apos", "imprimir", "mg", "pdf", "cep", "tal","ltda", "s/a", "jd" e "s.a". Além disso, foram removidos do texto números romanos e letras isoladas.

Uma vez convertidas as sentenças judiciais em sequências de *tokens*, passou-se à vetorização da base de dados, ou seja, à conversão das sentenças em vetores numéricos. Para tanto, foram utilizadas três abordagens distintas:

1. Modelagem de tópicos com *Randomized Truncated SVD* (Pedregosa et al., 2011; Halko et al., 2009) sobre vetores TF-IDF (Lane et al., 2019, p.70-96);
2. Modelagem de tópicos com *Latent Dirichlet Allocation* (Blei et al., 2003) sobre vetores BOW (*bag of words*);
3. Criação de sentence embeddings das sentenças judiciais a partir de *embeddings* pré-treinados.

Tais técnicas foram escolhidas por serem amplamente utilizadas em tarefas de processamento de linguagem natural (Lane et al., 2019). O objetivo de usar as três técnicas era testá-las, de modo a selecionar a que apresentasse os melhores resultados nas análises realizadas posteriormente.

No caso dos vetores TF-IDF e BOW, optou-se por uma abordagem de modelagem de tópicos (em lugar de utilizar diretamente os vetores) tendo em vista o tamanho do vocabulário, para evitar que as consequências da maldição da dimensionalidade (Lane et al., 2019, p.80) influenciassem nos resultados. Neste sentido, *Singular Value Decomposition* (Lane et al., 2019, p.123-128) e de *Latent Dirichlet Allocation* (Lane et al., 2019, p.134-137) são técnicas comumente utilizadas para extrair "tópicos" de dados textuais (com a consequente redução da dimensionalidade dos vetores por elas utilizados). Assim, ao invés de interpretar as sentenças como conjuntos de palavras, optou-se por representá-las de acordo com os tópicos nelas tratados, obtidos a partir dos referidos algoritmos.

A criação dos vetores TF-IDF e BOW foi feita a partir da base de dados convertida em *tokens*. No caso dos vetores TF-IDF, utilizou-se a implementação `TfidfVectorizer` da biblioteca scikit-learn (Pedregosa et al., 2011). Os vetores foram então fornecidos como entrada para o algoritmo `Randomized Truncated SVD`, que, ao final, reduziu a dimensão de cada sentença judicial para `1X800`. A variância total explicada obtida após a redução de dimensionalidade foi de $0,96$.

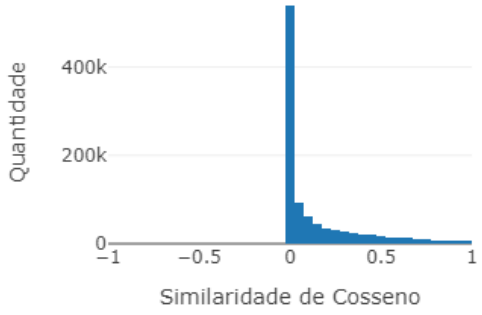
Já os vetores BOW foram criados com `CountVectorizer`, também disponível na biblioteca scikit-learn (Pedregosa et al., 2011). Efetuou-se um procedimento de *grid search* para definir o melhor número de componentes para o LDA, algoritmo que recebeu como entrada os vetores BOW. A métrica utilizada para esta finalidade foi a perplexidade, culminando, ao final, com a conversão de cada sentença em um vetor numérico de dimensão `1X35`.

Por fim, para criação dos *sentence embeddings*, cada sentença judicial foi segmentada em trechos menores com o uso da função `sent_tokenize` da biblioteca `NLTK` (Bird et al., 2009). A partir daí, com o uso de *embeddings* pré-treinados em português, cada trecho menor foi convertido em um *sentence embedding* de dimensão `1X768`, de modo que cada sentença judicial passou a ser representada como uma matriz de *sentence embeddings*. Ao final, tomou-se a média de todos os *embeddings*, de modo que cada sentença passasse a ser representada como um único vetor de dimensão `1X768`.

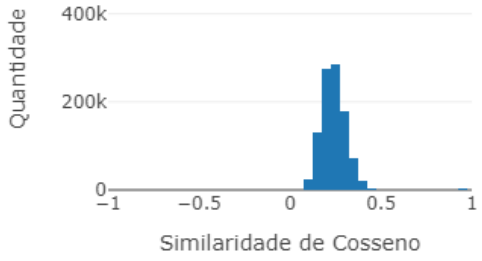
5.2 Similaridade de cosseno

Inicialmente, a fim de compreender o quão similares as representações numéricas das sentenças judiciais são entre si, computou-se a similaridade de cosseno (Manning et al., 2009, p. 120-122) de todos os pares de vetores para cada um dos três conjuntos de vetores (o primeiro, formado a partir de *Sentence Embeddings*, o segundo a partir de LDA e o terceiro a partir de TF-IDF). Na prática, computaram três matrizes de similaridade $S_{1422\times1422}$ nas quais cada elemento s_{ij} corresponde à similaridade de cosseno entre os vetores i e j. Assim, tem-se que s_{ij} pode assumir qualquer valor no intervalo $[-1, 1]$, sendo que, quanto mais próximo de 1, maior a similaridade entre os vetores i e j. As distribuições dos diferentes valores obtidos para cada um dos conjuntos de vetores são representadas nos histogramas constantes da Figura 5.1.[4]

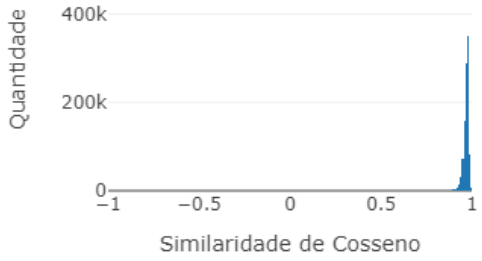
[4]Para cada uma das três matrizes, os diferentes valores de similaridade de cosseno foram extraídos da matriz de similaridade excluindo-se os valores constantes da diagonal principal e todos aqueles situados à sua direita, para evitar repetições.



(a) Similaridades de cosseno para LDA



(b) Similaridades de cosseno para TF-IDF



(c) Similaridades de cosseno para *Sentence Embeddings*

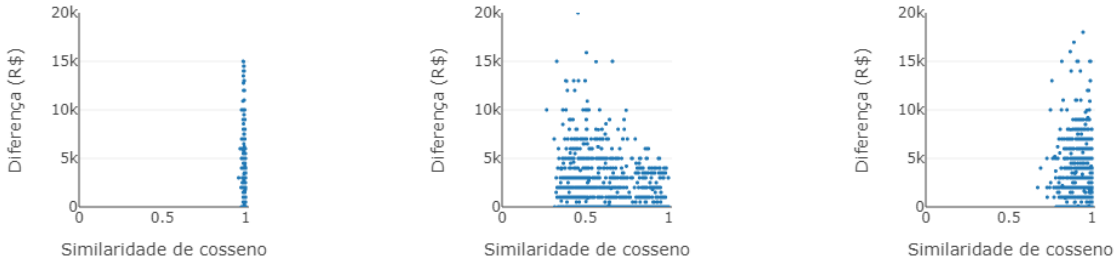
Figura 5.1. Valores do Índice de Davies-Bouldin ("DBI") e do Coeficiente de Silhueta ("SC") para cada algoritmo para cada grupo de vetores.

Para os vetores LDA e TF-IDF nota-se uma concentração maior de valores de similaridade de cosseno próximos ao zero, mas positivos. A baixa similaridade pode decorrer do fato de que os vetores têm, em geral, muitas dimensões, sendo que diferença na distribuição de valores entre os vetores LDA e TF-IDF possivelmente se explica pela maior dimensionalidade no segundo grupo (800 dimensões) em relação ao primeiro (35 dimensões) e pelo fato de que pode ter havido, no primeiro grupo, uma perda maior de informação ao construir vetores LDA a partir de vetores BOW. Quanto aos *Sentence Embeddings*, é possível conjecturar que a alta similaridade entre os textos decorre da forma como os vetores foram

construídos: como se partiu de *embeddings* pré-treinados a partir de uma base de dados genérica (e não especificamente jurídica), é possível que o fato de todos os textos terem origem no mundo jurídico fez com que os vetores ficassem muito parecidos entre si.

Na sequência, para cada linha das três matrizes S, identificou-se a coluna com o maior valor, correspondente ao vetor mais próximo àquele a que a linha se refere (excluindo ele mesmo), e, para cada par de sentenças mais similares entre si, computou-se o módulo das diferenças dos valores de indenização nelas fixados, a fim de entender se uma maior similaridade implicaria em uma menor diferença nos referidos valores (o que seria, idealmente, esperado).

Os resultados podem ser visualizados na Figura 5.2. Idealmente, o gráfico deveria ter um formato cônico, de modo que, quanto menores os valores no eixo y, maiores os verificados no eixo x, e vice-versa. No entanto, o que se nota é uma alta concentração de pares de vetores próximos ao valor 0 no eixo y que têm suas similaridades de cosseno fixadas entre $\approx 0,59$ e 1 (muito embora a maior concentração esteja mais próxima, de fato, de 1).



(a) Similaridade de cosseno por diferenças de valores para *Sentence Embeddings*.

(b) Similaridade de cosseno por diferenças de valores para vetores TF-IDF.

(c) Similaridade de cosseno por diferenças de valores para vetores LDA.

Figura 5.2. Diferenças de cosseno pelo módulo das diferenças nas indenizações.

Os achados evidenciam uma inexistência da esperada correlação entre semelhanças e valores, sobretudo para os vetores LDA e para os *Sentence Embeddings*. Uma discussão a este respeito será realizada no Capítulo 6.

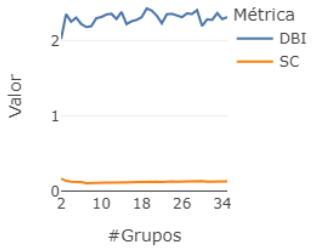
5.3 Agrupamento

Cada um dos três algoritmos de agrupamento testados (`K-Means`, `Agglomerative Clustering` e `Gaussian Mixture`) recebeu como entrada os vetores gerados na fase de pré-processamento descrita na Seção 5.1. A fim de investigar a capacidade de cada um destes algoritmos gerar grupos bem definidos e considerando a falta de rótulos que indiquem a que grupo cada sentença deve pertencer, os 3 algoritmos foram testados com o número de grupos variando de 2 até o número de diferentes valores de indenização.[5] As métricas utilizadas para a avaliação foram: Coeficiente de Silhueta (Rousseeuw, 1987) e Índice de Davies-Bouldin (Davies & Bouldin, 1979). Os resultados constam da Figura 5.3, que evidencia, para cada algoritmo e para cada conjunto de vetores, os valores dos referidos índices (representado no eixo `x`) de acordo com o número de grupos computados (representado no eixo `y`).

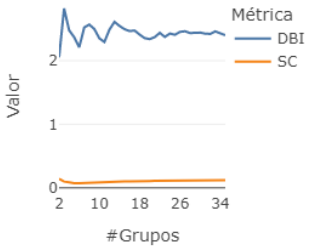
Os resultados obtidos indicam que os três algoritmos testados apresentaram dificuldades para agrupar as sentenças independentemente do método utilizado para convertê-las em vetores numéricos (*Sentence Embeddings*, LDA ou TF-IDF com `PCA`). Os valores baixos dos Coeficientes de Silhueta e distantes do zero de Davies-Bouldin sugerem grupos mal definidos em todos os casos - muito embora com

[5] É razoável esperar que sentenças similares entre si tenham valores fixados a título de indenização que sejam próximos, motivo pelo qual entendeu-se que, no máximo, o número de grupos deve ser igual ao número de diferentes sentenças.

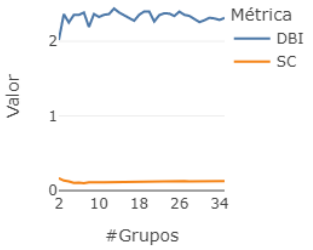
uma performance um pouco melhor para os vetores de entrada criados a partir de LDA -, o que está em consonância com os achados da Seção 5.2.



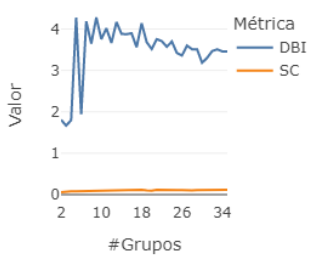
(a) K-Means com Embeddings



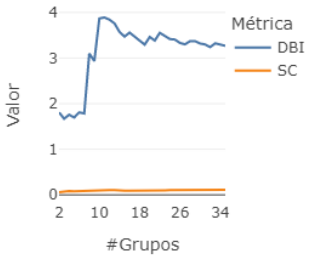
(b) Agglomerative Clustering com Embeddings



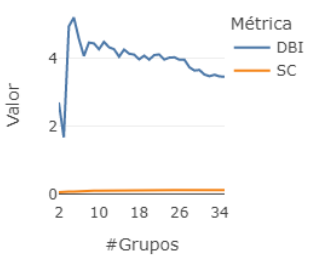
(c) Gaussian Mixture com Embeddings



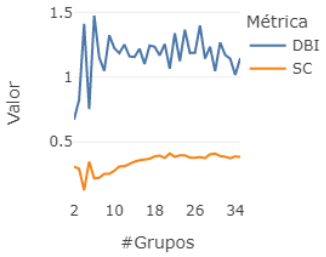
(d) K-Means com TF-IDF



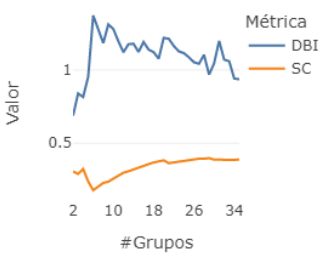
(e) Agglomerative Clustering com TF-IDF



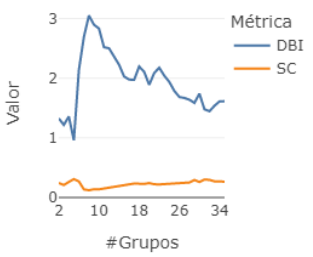
(f) Gaussian Mixture com TF-IDF



(g) K-Means com LDA



(h) Agglomerative Clustering com LDA



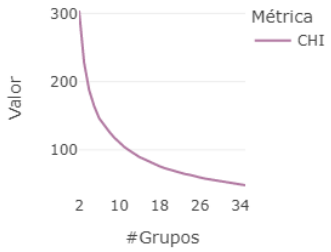
(i) Gaussian Mixture com LDA

Figura 5.3. Valores do Índice de Davies-Bouldin ("DBI") e do Coeficiente de Silhueta ("SC") para cada algoritmo para cada grupo de vetores.

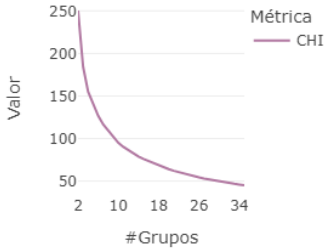
Aprofundando a análise, a fim de aferir o melhor número de grupos para cada algoritmo e para cada conjunto de vetores, utilizou-se, ainda, o Índice de Calinski-Harabasz (Caliński & JA, 1974), com os resultados expostos na Figura 5.4.

Considerando os valores calculados para Davies-Bouldin, Calinski-Harabasz e Coeficiente de Silhueta, definiram-se como ideais os números de grupos apontados na Tabela 5.1.

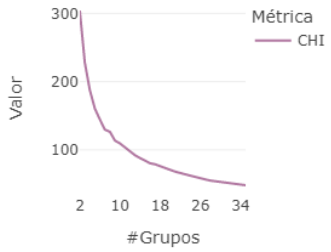
Os valores fixados a título de indenização, por grupo, varia de acordo com o evidenciado nos gráficos constantes da Figura 5.5. Avaliando os resultados (a partir dos gráficos e de uma análise individualizada de cada grupo), identificou-se que, independentemente dos vetores utilizados, todos os algoritmos colocaram, em um mesmo grupo, 79 sentenças judiciais proferidas por apenas seis magistrados, oriundos de três juízos distintos: "Unidade Jurisdicional - 2º JD da Comarca de Santa Luzia", "Unidade Jurisdicional Única - 1º JD da Comarca de Betim" e "Unidade Jurisdicional Única - 3º JD



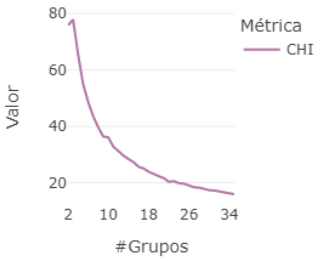
(a) K-Means com Embeddings



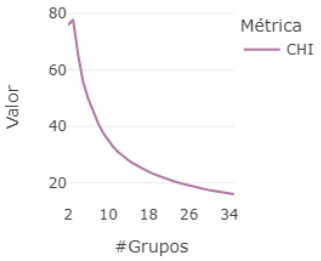
(b) Agglomerative Clustering com Embeddings



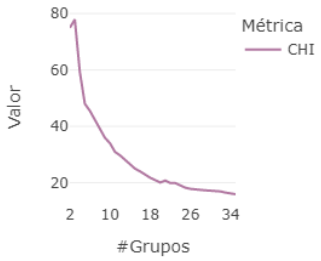
(c) Gaussian Mixture com Embeddings



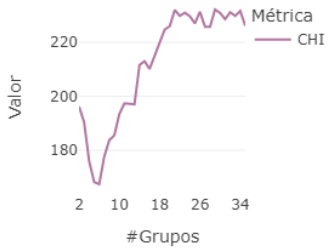
(d) K-Means com TF-IDF



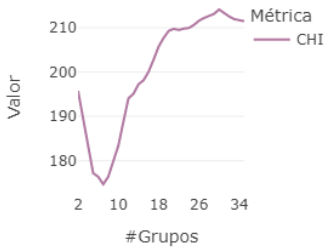
(e) Agglomerative Clustering com TF-IDF



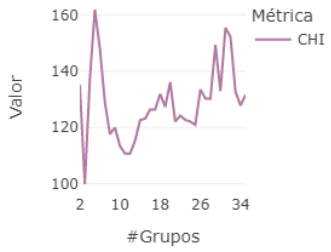
(f) Gaussian Mixture com TF-IDF



(g) K-Means com LDA



(h) Agglomerative Clustering com LDA



(i) Gaussian Mixture com LDA

Figura 5.4. Valores do Índice de Calinski-Harabasz ("CHI") para cada algoritmo para cada grupo de vetores.

da Comarca de Betim". No caso dos grupos criados a partir de *Sentence Embeddings*, os referidos 81 processos foram inseridos por todos os algoritmos em grupos grandes e mal definidos, que agregam uma série de outros processos.

A situação, porém, é mais interessante em se tratando dos demais vetores: os grupos `1`, `2` e `1` gerados respectivamente por `K-Means`, `Agglomerative Clustering` e `Gaussian Mixture` a partir de vetores TF-IDF com `PCA` contêm, cada um, os mesmos 82 processos, dos quais 79 são os mesmos mencionados anteriormente; e os grupos `1`, `1` e `2` gerados pelos mesmos algoritmos a partir de vetores LDA contêm, na ordem, 83, 84 e 83, sendo que os referidos 79 processos se repetem em todos.

Quanto aos valores fixados a título de indenização nos referidos processos, tem-se que em 55 foi definido o valor de R$ 6.000,00, em 22 o valor de R$ 2.000,00 e em cada uma das duas sentenças restantes os valores de R$ 4.000,00 e R$ 5.000,00.

Considerando tudo o que foi exposto, é possível inferir que os 79 processos agrupados em todos os casos são provavelmente muito similares entre si, o que foi objeto da investigação qualitativa descrita na Seção 5.3.1.

Tabela 5.1. Número de grupos selecionado para cada algoritmo para cada conjunto de vetores representativos de sentenças judiciais.

Entrada	*Algoritmo*	*#Grupos (n)*
TF-IDF	K-Means	3
	Agglomerative Clustering	3
	Gaussian Mixture	3
LDA	K-Means	2
	Agglomerative Clustering	2
	Gaussian Mixture	5
Embeddings	K-Means	2
	Agglomerative Clustering	2
	Gaussian Mixture	2

5.3.1 Análise qualitativa

Das 79 sentenças incluídas em um mesmo grupo por todos os algoritmos, independentemente dos vetores considerados, selecionaram-se onze: as seis primeiras foram escolhidas aleatoriamente entre aquelas nas quais fixou-se indenização por danos morais de R$ 6.000,00, as três seguintes entre as que houve fixação de R$ 2.000,00 e as duas últimas são aquelas nas quais as indenizações foram de R$ 4.000,00 e R$ 5.000,00. Dessa forma, todos os diferentes valores de indenização presentes naquele grupo de sentenças estão contemplados na amostra com pelo menos 10% de seus exemplares.

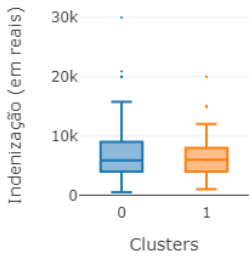
Os processos amostrados são todos muito similares entre si. Todos transcrevem, em suas respectivas fundamentações, a ementa do Agravo Regimental no Agravo de Instrumento de número 845.875/RN, proferido pela Quarta Turma do Superior Tribunal de Justiça, e a de uma decisão proferida pelo Tribunal de Justiça do Rio de Janeiro que não pode ser localizada pelo *site* do tribunal. A primeira assenta a tese de que o dano moral em caso de negativação indevida é presumido, enquanto a segunda estipula parâmetros para a fixação de danos morais, quais sejam: a necessidade de punir o responsável pela negativação indevida e a de compensar o dano sofrido pela vítima da negativação.

Todas as decisões constantes da amostra contêm diversos parágrafos idênticos, independentemente do valor fixado a título de danos morais. Além disso, os trechos que justificam o valor definido pelo juízo são bastante sucintos, muito embora exista uma pequena diferença quando o valor verificado foi R$ 6.000,00 ou R$ 2.000,00. Com efeito, no seguinte exemplo de trecho que fundamenta o valor de R$ 6.000,00 (idêntico para outras sentenças), lê-se:

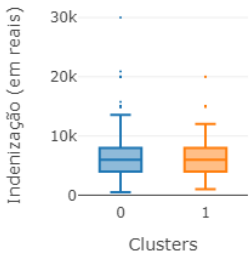
> *Assim, o valor da indenização decorrente do dano moral deve ser suficiente para reparar o dano do ofendido e servir como meio didático ao condenado para não reiterar a conduta ilícita. Lado outro, deve ser significativa, economicamente, para o causador do dano, mas não tão elevada de forma a consistir vantagem desmedida para o ofendido.*
>
> *Nesse diapasão, consideradas as peculiaridades do caso já abordadas e atento aos parâmetros do artigo 6º, da Lei 9.099, de 1995, entendo que o valor da indenização deva ser arbitrado em R$ 6.000,00 (seis mil reais), devendo, ainda, ser retirado o nome/CPF do autor do cadastro de inadimplentes.*

Já no que fundamenta o valor de R$ 2.000,00, lê-se:

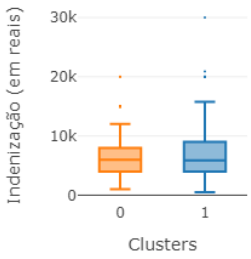
> *Assim, o valor da indenização decorrente do dano moral deve ser suficiente para reparar o dano do ofendido e servir como meio didático ao condenado para não reiterar a conduta ilícita. Lado outro, deve ser significativa, economicamente, para o causador do dano, mas não tão elevada de forma a consistir vantagem desmedida para o ofendido.*
>
> *Nesse diapasão, consideradas as peculiaridades do caso já abordadas e atento aos parâmetros do artigo 6º, da Lei 9.099, de 1995, entendo que o valor da indenização deva ser arbitrado em R$ 2.000,00 (dois mil reais),* ***ante a existência de outra negativação posterior em seu nome como consta na contestação da parte demandada****, devendo, ainda, ser retirado o nome/CPF do autor do cadastro de inadimplentes.*



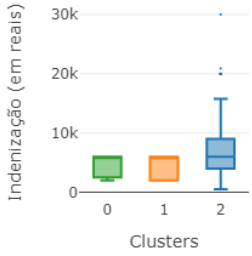
(a) K-Means com Embeddings



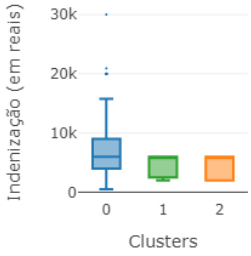
(b) Agglomerative Clustering com Embeddings



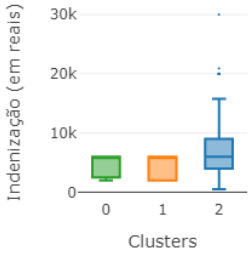
(c) Gaussian Mixture com Embeddings



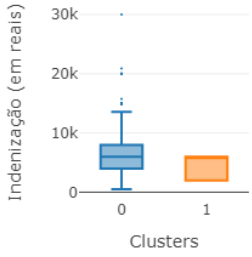
(d) K-Means com TF-IDF



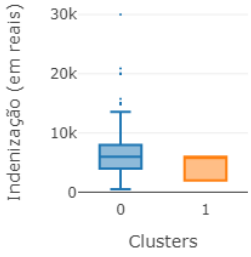
(e) Agglomerative Clustering com TF-IDF



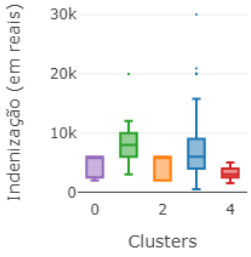
(f) Gaussian Mixture com TF-IDF



(g) K-Means com LDA



(h) Agglomerative Clustering com LDA



(i) Gaussian Mixture com LDA

Figura 5.5. Valores de indenização por grupo por algoritmo de agrupamento.

Como se vê, os trechos são essencialmente iguais, senão pelo trecho em negrito (que, aliás, já continha a mesma formatação no documento original). Nos demais julgados em que o valor da indenização é de R$ 2.000,00, o mesmo padrão se verifica, mas com pequenas alterações no excerto em destaque.

A análise da amostra evidencia, por um lado, um alto grau de similaridade entre os textos, que são todos praticamente iguais. Por outro, desvela um possível critério utilizado para fixação do valor da indenização por danos morais no âmbito do juízo "Unidade Jurisdicional Única - 1º JD da Comarca de Betim", qual seja: se existe outra negativação em nome do autor para além da discutida no processo, o valor fixado pelo juízo é menor.

Apesar disso, os mesmos dois parágrafos que fundamentam uma indenização de R$ 6.000,00 também foram utilizados para fundamentar as indenizações de R$ 5.000,00 e de R$ 4.000,00, sem que haja nos textos qualquer diferença senão pelo valor em si.

Quanto aos réus, tem-se que todos os processos nos quais foram fixados valores de R$ 6.000,00, R$ 2.000,00 e R$ 4.000,00 têm como ré a VIVO. Todas as sentenças, frise-se, fundamentam os valores de forma praticamente igual, de modo que, como a ré é a mesma em todos os casos, a variação - ao menos se

considerados apenas os dados disponíveis - existe somente quanto ao magistrado que proferiu a decisão e, no caso daquele responsável pelas sentenças de R$ 6.000,00 e R$ 2.000,00, quanto à existência ou não de prévia negativação do nome do autor.

Digno de nota, ainda, foi o fato de que cinco sentenças constantes da amostra tinham, ao final, a inscrição "<INSERIR NOME DO JUIZ>", permitindo inferir que houve a utilização de um modelo de sentença, o que justificaria o alto grau de similaridade verificado, mas não as diferenças verificadas nos valores R$ 6.000,00, R$ 5.000,00 e R$ 4.000,00, todos fixados a partir da mesma fundamentação.

Por fim, a fim de comparar o grupo de 79 sentenças que se repete em todos os agrupamentos com as demais sentenças constantes da base, gerou-se uma nova amostra aleatória, composta por quinze sentenças, dessa vez considerando a população das decisões proferidas por juízos distintos daqueles considerados nas amostragens anteriores.

Na última amostra, dez das quinze sentenças foram proferidas no âmbito da Justiça Comum, sendo todas oriundas de juízos distintos. Além disso, os valores fixados a título de indenização por danos morais foram: R$ 5000,00, com recorrência de quatro vezes; R$ 6000,00, com recorrência de três vezes; R$ 2000,00, R$ 2500,00, R$ 4000,00, R$ 12000,00 e R$ 8000,00, com recorrência de uma vez cada.

Nos trechos em que foram fixados os valores a título de indenização por danos morais, todas as sentenças têm o mesmo caráter genérico, havendo apenas uma que considerou o período pelo qual o nome do autor ficou negativado como critério específico para fixação do valor, muito embora o impacto do tempo sobre o montante não tenha ficado claro.

5.4 Regressão

Na sequência, passou-se à utilização de algoritmos de aprendizado de máquina para fins de análise dos dados. Ressalta-se que o presente estudo não tem o objetivo de desenvolver um algoritmo que seja extremamente eficiente para resolver um determinado problema: a ideia é recorrer a algoritmos de regressão para fins de compreender os motivos pelos quais os resultados são uns ou outros, de modo que, se o algoritmo performa bem ou mal, o foco deste estudo está em entender o motivo pelo qual isso acontece, não necessariamente em buscar fazer com que os resultados sejam melhores.

Para entender a performance dos modelos de regressão, utilizaram-se duas métricas: `R2` e `variância explicada` (Pedregosa et al., 2011).

Inicialmente, utilizou-se como linha de base um modelo de regressão linear. Além dele, foram treinados um `SVR` e um `SGD`, sendo que, no caso destes dois últimos, os valores de alguns parâmetros foram definidos a partir de `GridSearch` com validação cruzada. Os melhores resultados, obtidos para vetores LDA, estão na tabela 5.2.

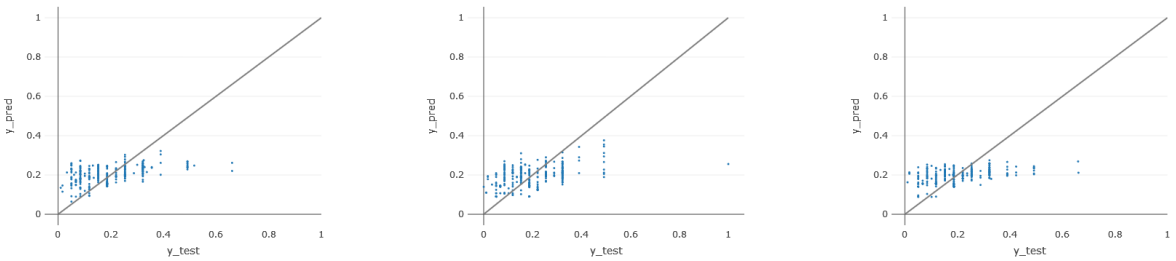
Tabela 5.2. Métricas de avaliação por modelo de regressão.

Vetores	*Algoritmo*	*Parâmetros*	*R2*	*Variância explicada*
	Linear	Padrão `sklearn`	0,206024	0,213607
LDA	SVR	`C`: 1 `coef0`: 2 `degree`: 3 `gamma`: 'scale' `kernel`: 'poly'	0,216437	0,220146
	SGD	`alpha`: 0.001 `learning_rate`: 'optimal' `loss`: 'huber' `penalty`: 'l2'	0,166456	0,172828

Os resultados, em geral, não são animadores, sendo que os motivos pelos quais isto acontece são discutidos em maiores detalhes no Capítulo 6. Uma das hipóteses é a de que os valores fixados a título de

indenização variam pouco, não são efetivamente contínuos e, em geral, são baixos, de modo que o modelo pode ter aprendido a prever valores menores independentemente da entrada (o que explica, aliás, valores de `R2 score` e de `variância explicada` próximos de zero na maior parte dos casos).

Os melhores resultados para cada algoritmo testado podem ser visualizados na Figura 5.6. O eixo x representa os valores de indenização (normalizados) que deveriam ter sido encontrados pelo modelo, enquanto, no eixo y, estão representados os valores que foram efetivamente encontrados. Um modelo perfeito deveria ter todos os seus pontos sobre a curva $x = y$ (plotada em cinza), de modo que, quanto maior a distância, pior a previsão.



(a) Regressão linear **(b)** SVR **(c)** SGD

Figura 5.6. Diferenças entre os valores esperado e encontrado de y para cada algoritmo de regressão.

Como se vê na Figura 5.6, para todos os modelos treinados, não há uma simetria em relação à curva $x = y$. Além disso, as maiores distâncias em relação àquela curva estão abaixo dela, sugerindo que há erros maiores quando o algoritmo recebe como entrada valores que sejam mais altos.

5.5 Classificação para identificação do valor fixado pelo juízo a título de indenização por danos morais

Ultrapassada a etapa de caracterização, passou-se à construção de um algoritmo que, dada uma sentença judicial proferida em primeira instância no âmbito dos juizados especiais ou da justiça comum, fosse capaz de identificar o valor fixado a título de danos morais pelo juízo.

Inicialmente, a biblioteca `NLTK` (Bird et al., 2009) foi utilizada para dividir as sentenças, no sentido jurídico, em sentenças menores, no sentido linguístico. Como resultado, cada sentença judicial foi subdividida, via de regra, em frases ou em conjuntos com poucas frases sequenciais. Então, para cada sentença judicial, identificaram-se as frases ou conjuntos de frases que continham menção a apenas um valor, fosse ele um valor irrelevante ou o valor fixado pelo juízo a título de indenização por danos morais. As frases e os rótulos dos valores nelas contidos foram armazenados em dois vetores X e y respectivamente, sendo que, na sequência, cada `string` armazenada no vetor X) foi submetida a um processo de lematização (o mesmo narrado na Seção 5.1), com a exclusão das palavras repetidas em mais de 99% e daquelas presentes em até 1% das `strings` em X. Ao final, restou um vetor X composto por uma listas de lemas, conforme exemplificado na Tabela 5.3.

Ressalta-se que as listas de `strings` resultantes do processo de lematização não continham os valores rotulados como `valor_fixado` ou como `valor_outros`, os quais foram excluídos a fim de evitar que o algoritmo aprendesse com base nos valores e não no texto propriamente dito. Tal medida é importante, tendo em vista que os valores fixados a título de indenização por danos morais, como já exposto anteriormente, não são contínuos e variam pouco, o que poderia tornar a tarefa de identificá-los muito simples para a base de dados utilizada para o treinamento, mas mais difícil se sobreviesse um novo valor nunca visto pelo algoritmo.

Na sequência, as listas de `strings` resultantes alimentaram um algoritmo TF-IDF (o mesmo mencionado na Seção 5.1), de modo que cada *string*, ao final, foi convertida em um vetor numérico de

Tabela 5.3. Exemplos de dados constantes do banco de dados utilizado para alimentar o classificador antes do pré-processamento.

X inicial	X após lematização	y
Relata a parte autora, em apertada síntese, que teve seu nome e CPF negativados indevidamente, em razão de débito oriundo de negócio jurídico não pactuado, no valor de R$ 31.345,10.	“negativar”, “negocio”, “juridico”, “razao”, “sintese”, “nao”, “debito”, “indevidamente”, “nome”, “autora”	`valor_outros`
Considerando o fundamento da demanda, em que o autor alega não ter celebrado o contrato nº 0204451779 que gerou a cobrança de R$214,49 e, consequentemente, a negativação, é da requerida o ônus de comprovar a origem e legitimidade do débito que ensejou a negativação.	“origem”, “demanda”, “gerar”, “onus”, “ensejar”, “requerida”, “nao”, “autor”, “cobranca”, “fundamento”, “negativacao”, “debito”, “contrato”, “considerar”, “alegar”, “comprovar”	`valor_outros`
Considerando-se as peculiaridades do caso, os intuitos ressarcitório e pedagógico da indenização moral, com vedação ao enriquecimento ilícito, vejo por bem arbitrar o quantum indenizatório em R$3.000,00, quantia esta que, com fulcro no artigo 6º da Lei nº 9.099, de 1995, entendo por justa e equânime a indenizar a parte lesada.	‘considerarse”, “pedagogico”, “equanime”, “fulcro”, “indenizar”, “arbitrar”, “justo”, “artigo”, “peculiaridade”, “quantia”, “indenizatorio”, “quantum”, “caso”, “entendo”, “ilicito”, “enriquecimento”, “indenizacao”, “moral”	`valor_fixado`
b) Condenar a parte ré a pagar ao autor, a título de danos morais, a quantia de R$ 7.000,00 (sete mil reais), valor esse devidamente corrigido a partir desta data até o efetivo pagamento, com base na Tabela da Corregedoria Geral de Justiça, acrescidos de juros de mora de 1% (um por cento) ao mês, nos termos do artigo 406 do Código Civil.	“termos”, “mora”, “justica”, “corregedoria”, “base”, “pagamento”, “efetivo”, “ate”, “data”, “devidamente”, “autor”, “civil”, “codigo”, “artigo”, “mes”, “tabela”, “quantia”, “re”, “pagar”, “condenar”, “juro”, “acrescer”, “corriger”, “moral”, “dano”, “titulo”	`valor_fixado`

dimensão 1×387.

Criou-se, então, uma matriz X de dimensão 3202×387, na qual cada linha continha os vetores que representavam uma frase ou um conjunto de frases de uma sentença judicial, e um vetor y de dimensão 3202×1, correspondente ao rótulo do valor contido naquela frase ou conjunto de frases na sentença original, o qual poderia ser `valor_outros` ou `valor_fixado`.

A matriz X e o vetor y, então, alimentaram um classificador `RandomForrest` (Breiman, 2001). As entradas foram subdivididas em cinco grupos distintos aleatórios, os quais foram novamente divididos em dois grupos, um de treino e um de teste. Os melhores resultados obtidos para um dos grupos são evidenciados na tabela 5.4.

Tabela 5.4. Avaliação dos resultados obtidos a partir do classificador RandomForrest.

	precision	*recall*	*f1-score*
`valor_fixado`	0,92	0,93	0,92
`valor_outros`	0,94	0,93	0,93
accuracy			0,93
macro avg	0,93	0,93	0,93
weighted avg	0,93	0,93	0,93

O modelo, como se depreende da Tabela 5.4, foi capaz de identificar corretamente os valores fixados a título de indenização por danos morais com uma precisão de 93%. As pequenas diferenças

entre os resultados obtidos para os rótulos `valor_fixado` e `valor_outros` provavelmente se explicam pelo fato de que a homogeneidade nos trechos em que há fixação de danos morais é maior do que nos demais, conforme evidenciado na Figura 5.7, no qual um mapa de árvore representa as ocorrências das cem palavras mais frequentes no grupo rotulado com `valor_fixado` (à esquerda da figura) e no grupo rotulado com `valor_outros` (à direita da figura). O tamanho dos retângulos é proporcional ao número de vezes em que aparece no *corpus* utilizado para a tarefa de classificação, e cores mais escuras representam ocorrências maiores, ao passo em que cores menos escuras fazem referência a um menor número de aparições.

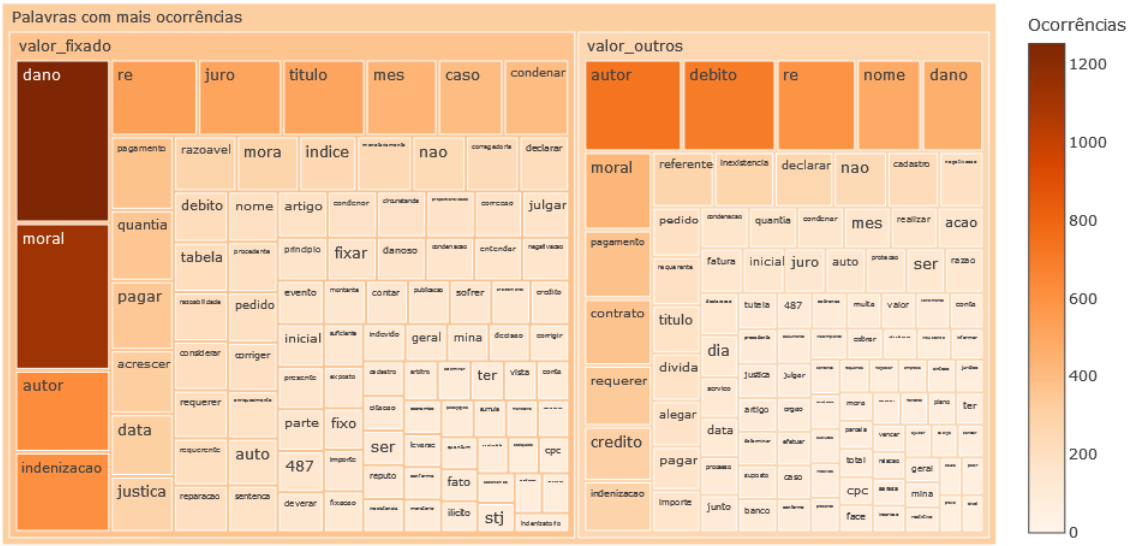


Figura 5.7. Mapa de árvore que evidencia o número de ocorrências de diferentes lemas em dois grupos: um composto por grupos de lemas extraídos de frases que continham o valor fixado pelo juízo a título de indenização por danos morais (à esquerda) e outro por grupos de lemas extraídos de frases que continham outros valores (à direita).

Como fica evidente na Figura 5.7, os retângulos no grupo composto por grupos de lemas rotulados com `valor_fixado` são em geral maiores e mais heterogêneos (quanto ao tamanho e quanto às cores), ao passo em que os retângulos do grupo `valor_outros` são menores e mais homogêneos. A figura dá conta de que, no primeiro grupo, há uma maior concentração de ocorrências de lemas específicos, como "moral" e "dano", que também aparecem no segundo grupo. Porém, no primeiro há uma grande ocorrência de outras lemas relativos a palavras que comumente aparecem, de forma específica, no dispositivo das sentenças judiciais, como é o caso de "condeno" (relativo a "condenar"), "juro" (relativo a "juros") e "título" (palavra esta que costuma integrar a expressão "a título"). A recorrência destas palavras em dispositivos de sentenças proferidas no âmbito de processos judiciais pode ser verificada qualitativamente - e faz sentido, considerando que o que o juiz faz ao dar provimento a um pedido de condenação a reparação por danos morais é, de fato, estabelecer uma condenação (daí o verbo "condenar"), consistente em uma obrigação de pagar um valor "a título de danos morais", sendo que este valor será acrescido de juros (daí o aparecimento do lema "juro").

Assim, acredita-se que os bons resultados obtidos com o classificador se devam à existência de uma alta homogeneidade nos trechos das sentenças judiciais em que há, de fato, a fixação do valor da condenação por danos morais (ao menos na base de dados utilizada neste estudo), tratando-se, portanto, de uma tarefa simples do ponto de vista computacional.

5.6 Breves considerações sobre a modelagem dos dados

As tarefas de tentar agrupar sentenças similares (partindo da premissa de que sentenças parecidas deveriam ter valores parecidos de indenização) e de tentar predizer os valores fixados a título de indenização com base exclusivamente nos textos das sentenças se mostraram infrutíferas do ponto de vista da eficiência dos algoritmos empregados para executar estas tarefas, mas permitiram que se fizessem algumas conjecturas a respeito dos resultados. Com efeito, talvez o valor fixado a título de indenização por danos morais seja um parâmetro ruim para definir o quão bons são os modelos produzidos a partir dos dados. Possivelmente, ainda, a tarefa de agrupamento restou prejudicada pela alta dimensionalidade dos dados na maior parte dos casos.

Em qualquer cenário, não há dúvidas de que o universo restrito de sentenças judiciais limitou as análises realizadas, tendo em vista que um corpo de 1.422 sentenças é, certamente, ínfimo diante do montante total efetivamente existente nos sistemas do TJMG.

A dificuldade na realização de uma correlação entre os textos das sentenças e os valores fixados, sobretudo considerando que as demandas judiciais objeto de estudo são absolutamente simples, pode ser um indicativo de que as sentenças judiciais não são tão bem fundamentadas quanto deveriam - talvez porque os parâmetros a serem considerados para fixação de danos morais são obscuros e altamente subjetivos. Com efeito, não há, em tese, nada que imponha o dever específico de o magistrado explicitar como cada um dos requisitos necessários para fixação do montante indenizatório foi considerado no caso concreto, o que faz com que a simples listagem dos requisitos seguida do valor seja suficiente, na prática, do ponto de vista processual, para fundamentá-lo.

Por outro lado, a tarefa de identificar os valores fixados a título de danos morais em uma sentença judicial se mostrou simples, o que pode servir como guia para esforços futuros de caracterização e de extração de informações a partir de bases de dados não estruturadas de sentenças judiciais.

Feitas estas digressões e concluídas as análises pretendidas, todas as hipóteses levantadas no início do estudo se encontram em condição de ser avaliadas, o que é feito em maiores detalhes no Capítulo 6.

Capítulo 6

Conclusão

A partir da metodologia e da análise dos resultados descritos nos Capítulos 3, 4 e 5, busca-se confirmar ou refutar as hipóteses que motivaram o presente estudo e, na sequência, apontar as limitações do presente estudo e possíveis direções para futuras investigações.

6.1 Ausência de padrão na fixação de reparações por danos morais

A hipótese de que é possível identificar um padrão bem estabelecido para a fixação de reparações por danos morais foi refutada em parte. Os modelos de agrupamento treinados a partir de vetores numéricos tiveram, em todos os cenários, dificuldade na identificação de grupos, à exceção de um ou dois pequenos grupos compostos basicamente por decisões oriundas de magistrados vinculados à Unidade Jurisdicional Única do Juizado Especial de Betim, permitindo a inferência de que os textos, em geral, não são facilmente distinguíveis entre si para a maior parte da base de dados.

Além disso, quando calculada as similaridades de cosseno (Seção 5.2) a partir do *corpus*, detectou-se que variações nas similaridades não refletem proporcionalmente em variações nos valores fixados a título de indenização por danos morais, contradizendo a expectativa razoável de que textos similares entre si teriam valores que também o fossem. No geral, o que se identificou foi que as similaridades de cosseno são em geral altas (entre pares mais de vetores mais próximos entre si), concentradas entre $0,75$ e 1, independentemente de as diferenças nos valores das indenizações serem altas ou baixas.

A análise qualitativa (Seção 5.3.1) identificou que, para a maior parte das sentenças amostradas, uma mesma fundamentação poderia justificar um ou outro valor fixado a título de indenização por danos morais, sem que houvesse, no texto da decisão, diferenças que justificassem a opção feita pelo magistrado em um ou em outro sentido.

Contribui, ainda, para confirmar a tese de que não se pode identificar um padrão bem definido na fixação de reparações por danos morais o fato de que os algoritmos de regressão utilizados para tentar prever os valores fixados a título de indenização com base unicamente no texto das sentenças performam mal, com valores de `R2` e de `variância explicada` positivos e mais próximos de 0 que de 1. Uma possível explicação para este fenômeno seria o fato de que talvez o modelo tenha desenvolvido a tendência de diminuir os valores de `y_pred`, o que, aliás, teria respaldo na maior concentração de valores em patamares mais baixos, conforme explicitado no Capítulo 4.

Apesar disso, os dados sugerem a existência de um grau maior de padronização nas decisões oriundas da comarca de Betim. Com efeito, todos os algoritmos de agrupamento utilizados foram capazes de agrupar processos julgados no rito dos juizados especiais cíveis no âmbito daquela comarca. Todos os modelos treinados juntaram um grupo de 79 processos em um mesmo grupo, sugerindo não apenas que exista um padrão, mas que ele é suficientemente forte para ser detectado a partir de nove métodos distintos.

Contribuem para reforçar o maior grau de padronização das decisões oriundas de Betim as informações evidenciadas da Figura 4.5, que mostra uma baixa variabilidade dos valores praticados, além de uma alta concentração em valores pequenos, considerando que a mediana equivale também ao maior valor fixado naquele contexto, de `R$ 6.000,00`, e é igual ao terceiro quartil. O primeiro quartil, vale dizer, está muito próximo da margem inferior. O mesmo padrão pode ser visualizado na Figura 5.5, mais especificamente quanto aos grupos `1`, `2` e `1` gerados respectivamente por `K-Means`, `Agglomerative Clustering` e `Gaussian Mixture` a partir de vetores TF-IDF e quanto aos grupos `1`, `1` e `2` gerados pelos mesmos algoritmos a partir de vetores LDA.

Além disso, os três primeiros magistrados evidenciados na Figura 4.6 pertencem à comarca de Betim, sendo suas curvas absolutamente similares no que diz respeito à fixação de indenização por danos morais (todas as medianas e terceiros quartis têm o mesmo valor, de `R$ 6.000,00`).

Assim, tem-se que, em síntese, que, a partir dos métodos utilizados no presente estudo, não foi possível identificar um padrão geral bem estabelecido para a fixação de reparações por danos morais na base analisada, à exceção dos processos oriundos da Unidade Jurisdicional Única da Comarca de Betim, os quais têm sentenças muito parecidas entre si cujos valores variam pouco.

Acredita-se que estes resultados podem ser motivados pelo fato de que, em que pese os magistrados tenham o dever de motivar suas decisões judiciais, a fixação de indenizações por danos morais acaba sendo uma tarefa extremamente subjetiva, motivo pelo qual a sentença se limita a cumprir uma formalidade legal, sem que de fato se descrevam os pormenores que levaram à fixação de uma ou de outra quantia. Em diversos casos estudados quando da análise qualitativa descrita na Seção 5.3.1, verificou-se que as sentenças limitam-se a repetir os critérios jurisprudenciais para a definição dos valores das reparações, sem que o peso de cada um seja efetiva e objetivamente avaliado, conduta esta que possibilita o cumprimento de um critério formal, mas que não permite conhecer o que de fato levou o magistrado a optar por um valor e não por outro.

É possível, ainda, que as diferenças nos valores sejam explicadas por fatores não evidenciados no textos das sentenças, como documentos juntados ao processo, percepção do magistrado durante audiências, entre outros. A sentença, por si só, não capta estas influências, o que demandaria estudos de outras áreas do conhecimento.

6.2 Comparação entre valores praticados na justiça comum e nos juizados especiais

A partir da comparação entre valores fixados na justiça comum e nos juizados especiais a título de reparação por danos morais, no contexto específico da base de dados tratada neste estudo, foi possível confirmar a hipótese, formulada com fundamento no senso comum, de que os praticados nas varas cíveis são superiores aos verificados nos juizados.

A tendência se verificou não apenas na base de dados como um todo (Figuras 4.1 e 4.2), mas, em específico, também nas três comarcas nela mais representadas (Figura 4.5).

Em que pese este achado não possa ser extrapolado para o Poder Judiciário como um todo nem mais especificamente ao TJMG, os resultados dão uma sinalização no sentido de que o senso comum pode estar, efetivamente, correto, de modo que, havendo possibilidade de se escolher entre uma ou outra jurisdição, levando-se em consideração estritamente o possível ganho financeiro, pode ser mais vantajoso mover uma demanda de reparação por danos morais em virtude de negativação indevida na justiça comum que nos juizados especiais. Apesar disso, tal critério não é, por si só, o mais adequado, tendo em vista que a duração do processo também pode ser relevante no caso concreto, a depender dos interesses da parte promovente. Isso porque, em que pese os valores sejam, em geral, maiores em processos que tramitaram ou que tramitam na justiça comum, fato é que o tempo decorrido entre a distribuição da ação e a juntada da sentença também o são, conforme evidenciado pelas Figuras 4.10 e 4.12. Com efeito, a probabilidade de um processo terminar antes na Justiça Comum é menor que no Juizado Especial, conforme evidenciado pela comparação entre as curvas de sobrevivência dos dois cenários (Figura 4.13).

6.3 Utilização de técnicas de NLP e de aprendizado de máquina para análise de decisões judiciais

O estudo confirmou a hipótese de que é possível analisar decisões judiciais a partir da utilização de técnicas de `NLP` (pré-processamento e vetorização de texto) e de aprendizado de máquina (agrupamento e *regressão*).

Ainda que os resultados obtidos pelos algoritmos de regressão e de agrupamento não sejam considerados bons para tarefas preditivas, o objetivo, aqui, não é predizer valores, mas entender as métricas e utilizar os resultados para enriquecer a compreensão das bases de dados e dos motivos que os levam a ser como são.

Os resultados obtidos para os algoritmos de regressão linear podem ser explicados, acredita-se, pela baixa variabilidade nos valores fixados a título de indenização por danos morais, concentrados em montantes mais baixos que mais altos. Acredita-se que, por causa destes fatores, os algoritmos acabaram por aprender a atribuir valores baixos de indenização independentemente da entrada. É possível, porém, que os resultados se expliquem pelo pequeno tamanho da base de dados ou pela alta representatividade que alguns poucos juízos ou que alguns poucos magistrados nela têm, de modo que uma base maior e mais diversa poderia permitir que resultados mais interessantes fossem obtidos.

Vale ressaltar, neste ponto, que talvez a diferença entre os valores de indenização não sejam um bom *proxy* para a similaridade entre diferentes sentenças, de modo que a tarefa de tentar predizer valores a partir de textos seja inadequada, ainda que, idealmente, se pudesse imaginar uma relação necessária entre as duas variáveis. É possível que, se outras informações (como aquelas constantes de documentos do processo) fossem dadas como entrada para os algoritmos de regressão aqui considerados, ou mesmo para outros, os resultados seriam diferentes.

Além disso, considerando a baixa variabilidade nos valores de indenização, talvez o problema de regressão fosse melhor enfrentado sob uma perspectiva de classificação, o que não se fez em virtude do desbalanceamento existente nos diferentes valores da entrada. Tentou-se rotulá-los de acordo com diferentes percentis, mas, em todos os casos, as classes resultantes ficaram desbalanceadas, com um número maior de elementos em classes que correspondiam a faixas de valor menores. Uma abordagem possível, não testada, seria a criação de dados similares aos constantes da base para fins de balanceamento entre as diferentes classes, mas tal expediente não foi adotado neste estudo.

Na mesma linha, a dificuldade de se encontrarem vários grupos bem definidos de sentenças também pode ser explicada pela alta homogeneidade das decisões ou pela falta de mais processos na base de dados, com uma maior representatividade de outros juízos e magistrados.

Por outro lado, a possibilidade de se encontrar um grupo bem definido de sentenças pode ser explicada pelo alto grau de semelhança das decisões constantes da base de dados que foram proferidas no âmbito do Juizado Especial Cível de Betim, evidenciado pelos resultados da análise qualitativa (Seção 5.3.1.

6.4 Criação de modelo para identificação automatizada de valores de indenização

A identificação dos valores fixados a título de indenização por danos morais nas sentenças constantes da base de dados se mostrou possível a partir da utilização de um algoritmo simples e rápido, com boa precisão. Os trechos em que tais valores aparecem são, em geral, muito similares entre si, contendo palavras como “condeno”, “indenização”, “danos” e “morais”, o que provavelmente explica os bons resultados. Com efeito, se a homogeneidade pode prejudicar a tarefa de predizer valores ou de agrupar sentenças, passa a ser uma boa característica quando se pretende identificar trechos similares em textos diferentes.

O método para identificação de valores em dados não estruturados aqui proposto representa um possível avanço em relação ao apresentado por Nunes & Trenceti (2015), na medida em que permite estimar o erro na tarefa de classificação de forma mais objetiva, possibilitando que este fator seja considerado

para fins de se tirar conclusões mais assertivas a partir da análise dos valores.

6.5 Limitações e próximos passos

Apesar das dificuldades verificadas no Brasil para a realização de estudos com dados de processos judiciais (Colombo et al., 2017, p. 12), coletá-los, processá-los e analisá-los é possível, ainda que não seja necessariamente fácil fazê-lo sem algum grau de cooperação dos tribunais estudados. Com efeito, o presente estudo só foi possível a partir do fornecimento, pelo TJMG, de uma lista contendo uma série de informações sobre processos judiciais, o qual foi precedido de uma autorização de acesso ao sistema de recuperação da informação adotado internamente no tribunal. Esta foi uma das principais limitações do presente estudo, na medida em que as análises tiveram de se restringir ao pequeno espaço amostral de processos cujos números e demais informações foram fornecidos pelo tribunal.

Outro problema importante que teve de ser enfrentado neste estudo consiste no fato encontrar processos que se amoldem especificamente a um determinado assunto para fins de se analisá-lo não é, necessariamente, uma tarefa simples, à medida em que erros na identificação de assuntos são bastante recorrentes. Além disso, a tabela fornecido pelo TJMG continha informações desencontradas quanto à data de julgamento das demandas e não separava as partes do processo entre autores e réus.

Apesar disso, acredita-se que o presente trabalho pode contribuir para o avanço da jurimetria do Brasil, na medida em que busca enfrentar o problema de se realizar estudos jurimétricos a partir de dados não estruturados, abordagem que não é tão comum na literatura nacional. A metodologia proposta compreende as seguintes etapas, as quais podem ou não ser sequenciais como aqui se apresentam:

1. Definição do objeto de análise, das hipóteses de pesquisa e dos requisitos necessários para sua realização;

2. Obtenção dos dados, seja por meio de requisição aos órgãos objeto do estudo, seja por meios automatizados (aqui, se realizou por meio de um *web crawler*), seja por uma conjugação de ambos;

3. Pré-processamento com transformação dos dados textuais em representações numéricas (como vetores `TF-IDF` com dimensionalidade reduzida a partir de `PCA`, vetores `BOW` com dimensionalidade reduzida a partir de `LDA` ou *sentence embeddings*);

4. Processamento dos dados a partir de métodos puramente estatísticos e do uso de técnicas de aprendizado de máquina;

5. Análise de todos os resultados e das métricas utilizadas para avaliar os modelos de aprendizado de máquina, confrontando-os.

Em relação à vetorização de dados textuais, um futuro caminho interessante de pesquisa seria o desenvolvimento de *word embeddings* pré-treinados em bases de dados jurídicas, o que poderia contribuir sobremaneira para tarefas que envolvam o processamento de textos jurídicos, aproximando-as do estado da arte em matéria de processamento de linguagem natural.

Além disso, outro caminho de pesquisa que pode ser interessante é o aprofundamento da que aqui se apresentou, com uma base de dados maior que englobe, também, dados de outros tribunais brasileiros, de modo que seja possível realizar estudos comparativos para entender, a nível nacional, como está a prática de fixação de danos morais em demandas de negativação indevida. Podem ser utilizados, ainda, outros algoritmos que aqui não foram utilizados, a fim de entender se e como contribuem para a compreensão dos dados.

Por fim, considerando a rápida evolução, nos últimos anos, de *Large Language Models* capazes de executar tarefas impressionantes de geração de textos, ainda faltam pesquisas no sentido de investigar as possibilidades de adoção de técnicas de transferência de aprendizado para fazer com que possam ser utilizados com segurança no mundo jurídico. Além disso, em que pese a doutrina jurídica venha se ocupando de tratar dos impactos dos *Large Language Models* no Direito, faltam investigações que possibilitem a criação de métodos para aplicação destas tecnologias, por exemplo, no cotidiano forense, bem como que permitam a avaliação de sua eficiência.

Referências

Alexandre Freitas Câmara (2022). *O Novo Processo Civil Brasileiro*. Grupo GEN, Barueri, 8 edição. ISBN 9786559772575.

Associação Brasileira de Jurimetria (2017). *Os maiores litigantes em ações consumeristas: mapeamento e proposições*. Justiça Pesquisa. CNJ, Brasília.

Baade, H. W. (1963). Foreword. *Law and Contemporary Problems*, 28:1–4.

Bird, S.; Klein, E. & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc. ISBN 978-0-596-51649-9.

Blei, D. M.; Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022. ISSN 1532-4435.

Brasil (1973). Lei nº 5.869, de 11 de janeiro de 1973. Institui o Código de Processo Civil. *Diário Oficial da República Federativa do Brasil*.

Brasil (1988). Constituição da República Federativa do Brasil de 1988.

Brasil (2014). Superior Tribunal de Justiça. Recurso Especial nº 1374284/MG. Relator: Min. Luis Felipe Salomão. *Diário da Justiça Eletrônico*.

Brasil (2016). Jurisprudência em Teses. Relatório técnico 59, Superior Tribunal de Justica, Brasília.

Brasil (2021). Justiça em Números 2021. Relatório técnico, Conselho Nacional de Justiça, Brasília.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5--32.

Britto, L. M. T.; Lacerda, L. R. & Karninke, T. M. (2018). A crise do congestionamento do Poder Judiciário e a ingerência dos conflitos de massa no prejuízo do acesso à justiça. Seriam as técnicas coletivas de repercussão individual instrumentos necessários para desestimular a litigância habitual? In *Anais*, pp. 222--235, Vitória.

Bzdok, D.; Altman, N. & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15(4):233--234. ISSN 1548-7105.

Caliński, T. & JA, H. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods*, 3:1--27.

Colombo, B.; Buck, P. & Miana, V. (2017). Challenges When Using Jurimetrics in Brazil—A Survey of Courts. *Future Internet*, 9:68.

Conselho Nacional de Justiça (2022). DATAJUD.

Couto, M. B. & Oliveira, S. P. d. (2016). Gestão da Justiça e do conhecimento: a contribuição da jurimetria para a administração da justiça. *Revista Jurídica - UNICURITIBA*, 2(43):771--801. ISSN 2316-753X.

Davies, D. L. & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224--227.

Diniz, B. S.; Amâncio, J. A.; Borges, M. R.; Cota, T. T. & Faria, R. M. (2020). RADAR: Uma contribuição da tecnologia da informação para a gestão de processos repetitivos no Tribunal de Justiça de Minas Gerais. *Revista de Precedentes Qualificados Tribunal de Justiça de Minas Gerais*, 2(2).

Gargano, R. d. S. & Nader, C. C. F. d. C. (2018). As controvérsias acerca da aplicação da Jurimetria da pena nas relações de consumo dos Juízados Especiais Cíveis. *Alumni - Revista discente da UNIABEU*, 6(11). ISSN 23183985.

Halko, N.; Martinsson, P.-G. & Tropp, J. A. (2009). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.

Heartex, Inc. (2022). Label Studio.

Lane, H.; Howard, C. & Hapke, H. M. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning, Shelter Island. ISBN 978-1-61729-463-1.

Leonardo Netto Parentoni (2012). *Reconsideração da personalidade jurídica: estudo dogmático sobre a aplicação abusiva da disregard doctrine com análise empírica da jurisprudência brasileira*. Tese de doutorado, Faculdade de Direito da Universidade de São Paulo, São Paulo.

Loevinger, L. (1949). Jurimetrics: The Next Step Forward. *Minnesota Law Review*, 33:455.

Loevinger, L. (1963). Jurimetrics: The Methodology of Legal Inquiry. *Law and Contemporary Problems*, 28:5–35.

Magalhães, M. N. & Lima, A. C. P. d. (2015). *Noções de probabilidade e estatística*. Editora da Universidade de São Paulo, São Paulo, 7 edição. ISBN 978-85-314-0677-5.

Maia, M. & Bezerra, C. A. (2020). Análise bibliométrica dos artigos científicos de jurimetria publicados no Brasil. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, 18:e020018. ISSN 1678-765X.

Manning, C. D.; Raghavan, P. & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press. ISBN 0-521-86571-9.

Moisés, J. ; Verissimo, J.; Corrêa, F.; Trecenti, J. A. Z.; Werner, G.; Nunes, E. & Nunes, M. G. (2019). Justiça Criminal, Impunidade e Prescrição. Relatório técnico, Conselho Nacional de Justiça, Brasília.

Moraes, R. (2017). Que tipo de saber é o Direito? Entre a ciência, a prudência e a técnica. *Revista da Faculdade de Direito UFPR*, 62(1):83--111. ISSN 2236-7284.

Muthukadan, B. (2018). Selenium with Python.

Novaes, R. V.; Magalhães Gomes, M. F. d. & Valentini, R. S. (2018). Desenvolvimento tecnológico e o futuro da atividade jurídica. In *Desenvolvimento tecnológico e o futuro da atividade jurídica*, volume 1, pp. 989--1007. Editora D'Plácido, Belo Horizonte.

Nunes, D. & Duarte, F. A. (2021). Jurimetria, Tecnologia e Direito Processual. In *Inteligência Artificial e Direito Processual: Os Impactos da Virada Tecnológica no Direito Processual*, p. 928. JusPodivm, Salvador.

Nunes, M. G. (2015). Tempo dos processos relacionados à adoção no Brasil: uma análise sobre os impactos da atuação do Poder Judiciário. Relatório técnico, Conselho Nacional de Justiça, Brasília.

Nunes, M. G. (2016a). Estudo sobre Varas Empresariais na Comarca de São Paulo. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Nunes, M. G. (2016b). *Jurimetria: como a estatística pode reinventar o Direito*. Revista dos Tribunais, São Paulo.

Nunes, M. G. (2020). Estudo jurimétrico sobre Execução de Contratos: Relatório Doing Business. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Nunes, M. G. & Berger, R. (2020). Observatório do Mercado de Capitais: Atividade Disciplinar da CVM. Relatório técnico, Associação Brasileira de Jurimetria e Associação Brasileira das Companhias Abertas, São Paulo.

Nunes, M. G.; Corrêa, F.; Trecenti, J. A. Z. & Jesus Filho, J. d. (2019). Avaliação do Impacto de Critérios Objetivos na Distinção Entre Posse para Uso e Posse para Tráfico: Um estudo Jurimétrico. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Nunes, M. G.; Langeani, B.; Pollachi, N.; Trecenti, J. A. Z. & Corrêa, F. (2016). O Processamento de Homicídios no Brasil e a Estratégia Nacional de Justiça e Segurança Pública em três estados: Alagoas, Santa Catarina e São Paulo. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Nunes, M. G.; Trecenti, J. A. Z.; Cesario, P. & Roquim, P. (2014). CARF: Uma análise do sistema tributário. Relatório técnico.

Nunes, M. G.; Trecenti, J. A. Z.; Corrêa, F.; Coelho, F. U. & Stern, R. B. (2018). Os maiores litigantes em ações consumeristas: mapeamento e proposições. Relatório técnico, Conselho Nacional de Justiça, Brasília.

Nunes, M. G. & Trenceti, J. A. Z. (2015). Reformas de decisão nas câmaras de direito criminal em São Paulo. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Nunes, M. G.; Waisberg, I.; Sacramone, M. B.; Bumachar, J. & Trecenti, J. A. Z. (2022). Observatório da Insolvência: Processos de Recuperação Judicial no Rio de Janeiro. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Ozturkmenoglu, O. & Alpkocak, A. (2012). Comparison of different lemmatization approaches for information retrieval on turkish text collection. In *2012 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 1–5.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825--2830.

Pereira, C. M. d. S. (2022). *Responsabilidade Civil*. Forense, Rio de Janeiro, 13 edição. ISBN 9786559644926.

Pilatto, A. E. & Schumak Melo, F. (2020). Contra dados não há argumentos: teoria pura do direito e jurimetria. 2(1). ISSN 2674-9386.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. ISSN 0377-0427.

Schwengber, S. B. (2006). *Mensurando a eficiência no sistema judiciário: métodos paramétricos e não-paramétricos*. Tese (Doutorado em Economia), Universidade de Brasília, Brasília.

Serra, M. M. P. (2013). Como utilizar elementos da estatística descritiva na jurimetria. *Revista Eletrônica do Curso de Direito das Faculdades OPET*, 4(10). ISSN 2175-7119.

Unger, A. J.; Neto, J. F. d. S.; Fantinato, M.; Peres, S. M.; Trecenti, J. A. Z. & Hirota, R. (2021). *Process Mining-Enabled Jurimetrics: Analysis of a Brazilian Court's Judicial Performance in the Business Law Processing*, p. 240–244. Association for Computing Machinery, New York, NY, USA.

Waisberg, I.; Sacramone, M. B.; Nunes, M. G. & Corrêa, F. (2016). Recuperação Judicial nas Varas da Capital. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Waisberg, I.; Sacramone, M. B.; Nunes, M. G.; Corrêa, F. & Trecenti, J. A. Z. (2022a). Observatório da Insolvência. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Waisberg, I.; Sacramone, M. B.; Nunes, M. G. & Trecenti, J. A. Z. (2022b). Observatório da Insolvência - Fase 3: Falências no Estado de São Paulo. Relatório técnico, Associação Brasileira de Jurimetria, São Paulo.

Zabala, F. J. & Silveira, F. F. (2014). Jurimetria: Estatística aplicada ao Direito. *Revista Direito e Liberdade*, 16(1):87--103. ISSN 2177-1758.