UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS

UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO E SISTEMAS

NÍVEL DOUTORADO

LÚCIA ADRIANA DOS SANTOS GRUGINSKIE

TEMPO DE ATRAVESSAMENTO DE AÇÕES CÍVEIS NA JUSTIÇA FEDERAL DE 2^{O} GRAU: AJUSTE DE MODELOS BASEADOS EM REDES NEURAIS E MÁQUINA DE VETOR SUPORTE

SÃO LEOPOLDO

2018

Lúcia Adriana dos Santos Gruginskie

Tempo de atravessamento de ações cíveis na Justiça Federal de 2^o grau: ajuste de modelos baseados em redes neurais e máquina de vetor suporte

Tese apresentada como requisito parcial para obtenção do título de Doutor em Engenharia de Produção e Sistemas, pelo Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Universidade do Vale do Rio dos Sinos - UNISINOS.

Orientador: Prof. Dr. Miguel Afonso Sellitto

São Leopoldo

2018

G885t Gruginskie, Lúcia Adriana dos Santos
Tempo de atravessamento de ações cíveis na Justiça Federal de 2º grau : ajuste de modelos baseados em redes neurais e máquina de vetor suporte / por Lúcia Adriana dos Santos Gruginskie. – 2018.
209 f. : il. ; 30 cm.

Tese (doutorado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Engenharia de Produção e Sistemas, São Leopoldo, RS, 2018.
"Orientador: Dr. Miguel Afonso Sellitto".

1. Máquina de vetor suporte. 2. Redes neurais. 3. Análise de sobrevivência. 4. Tempo de atravessamento processual. 5. Gestão pública. I. Título.

CDU: 658.5:347.932

Catalogação na Publicação (CIP):
Bibliotecário Alessandro Dietrich - CRB 10/2338

Lúcia Adriana dos Santos Gruginskie

Tempo de atravessamento de ações cíveis na Justiça Federal de 2^o grau: ajuste de modelos baseados em redes neurais e máquina de vetor suporte

Tese apresentada como requisito parcial para obtenção do título de Doutor em Engenharia de Produção e Sistemas, pelo Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Universidade do Vale do Rio dos Sinos - UNISINOS.

Aprovado em 03 de Setembro de 2018.

BANCA EXAMINADORA:

Prof. Dra. Miriam Borchardt – Unisinos
Avaliador

Prof. Dra. Vivian Sebben Adami – Unisinos
Avaliador

Prof. Dr. Leonardo Dagnino Chiwiacowsky –
UCS Avaliador Externo

Prof. Dr. Gabriel Vidor – UCS
Avaliador Externo

Prof. Dr. Miguel Afonso Sellitto (Orientador)

Visto e permitida a impressão
São Leopoldo

Prof. Dr. Luiz Alberto Oliveira Rocha
Coordenador PPG em Engenharia de Produção
e Sistemas

À Hanna

RESUMO

Os órgãos do Poder Judiciário, como exemplo de instituições públicas, são fundamentais para o desenvolvimento econômico e social. Porém, os principais problemas do judiciário brasileiro, apontados pelo Ministério da Justiça, são o alto número de processos em estoque, a falta de acesso à justiça e a morosidade, considerada como o principal aspecto da crise do judiciário. Neste sentido, esta tese propõe estruturar e comparar modelos de previsão de tempo de atravessamento processual de ações judiciais civis na justiça federal de 2^a Instância, para servir de informação para as partes processuais e administração. O estudo foi realizado no Tribunal Regional Federal da 4^a Região, com dados de processos cíveis baixados em 2017. Para tanto, foram comparados quatro modelos para o tempo de atravessamento. O primeiro modelo foi ajustado através de redes neurais para regressão com o uso do algoritmo retroalimentação, o segundo utilizou máquina de vetor suporte para regressão, através da biblioteca Libsvm. O desempenho destes dois modelos, calculados pela medida RMSE, foi comparado ao desempenho da aplicação de análise de sobrevivência, o terceiro modelo, considerada técnica habitual para análise de estudos quantitativos de tempo. A variável resposta usada foi o tempo em dias entre a autuação do processo no 2^o Grau e a baixa, escolhida entre indicadores usados em trabalhos acadêmicos e por órgãos judiciais do Brasil e da Europa. O quarto modelo foi ajustado utilizando máquina de vetor suporte para classificação, através da biblioteca LIBSVM. A variável resposta para o ajuste deste modelo foi transformada em ordinal por meio da estratificação em faixas de tempo, o que permitiu o cálculo da medida acurácia para aferir o desempenho. As covariáveis usadas para os ajustes foram categóricas e estavam disponíveis no banco de dados do TRF. Após os ajustes, foram aplicadas regras de associação às faixas de tempo com o objetivo de encontrar as características dos processos mais lentos e morosos. Também foi analisada a viabilidade de estabelecer parâmetros de tempos razoáveis. Para utilização em previsões, sugere-se o modelo de classificação de faixa de tempo e para o estabelecimento de padrões, ou o modelo ajustado por redes neurais ou por máquina de vetor suporte. Entre as sugestões para trabalhos futuros estão a construção de uma tábua de baixa de processos, análoga às tábuas atuariais de mortalidade, e o estabelecimento de padrões para considerar processos como morosos.

Palavras-chaves: Máquina de vetor suporte. Redes neurais. Análise de sobrevivência. Tempo de atravessamento processual. Gestão pública.

ABSTRACT

The Courts as an example of public institutions, are fundamental for the economic and social development. However, the main problems of the Brazilian judiciary, pointed out by the Ministry of Justice, are the high number of cases in stock, lack of access to justice and slowness, considered as the main aspect of the crisis of the judiciary. In this sense, this thesis proposes to structure and compare models of lead time of civil lawsuits in the federal court of second instance, to serve as information for the parties of lawsuit and the administration. The study was conducted at the Tribunal Regional Federal da 4^a Região, with data from civil lawsuits termined in 2017. Four models for lead time lawsuits were compared. The first was fitted by neural network for regression with the backpropagation algorithm; the second model was fitted by using support vector machine for regression with the Libsvm library. The performance of these two models, calculated by the RMSE measurement, was compared to the performance of the survival analysis model, considered as the usual technique for the analysis of quantitative time studies. The dependent variable used was the time in days between the arraignment date and the case disposition date, chosen among indicators used in academic studies and by judicial Courts of Brazil and Europe. The fourth model was fitted using vector support machine for classification, using the Libsvm algorithm. The dependent variable was transformed into ordinal by means of the stratification in time bands, which allowed the calculation of the measurement as accuracy and precision. The independent variables were categorical and were available in the TRF database. After, rules of association were applied to the time bands in order to find the characteristics of the most frequent band and the more time consuming lawsuits. The feasibility of establishing parameters of reasonable times was also analyzed. The time-band classification model is suggest to use in forecasts and it can use the model adjusted by neural networks or by the support vector machine in use to establish standards of time. Among the suggestions for future work are the construction of a life tables of lawsuits, analogous to the actuarial tables, and the establishment of standards to consider reasonable lead time.

Key-words: Support vector machine. Neural network. Survival analysis. Lawsuit lead time. Public management.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BSC	*Balanced Scorecard*
CEPEJ	The European Comission for the Efficiency of Justice
CNJ	Conselho Nacional de Justiça
CPC	Código de Processo Civil
CRISP	*Cross-Industry Standard Process for Data Mining*
Gespública	Programa Nacional de Gestão Pública e Desburocratização
JE	Justiça Estadual
JF	Justiça Federal
LIBSVM	Biblioteca de máquina de vetor suporte
MEGP	Modelo de Excelência em Gestão Pública
RMSE	Raiz quadrada média do erro de predição ou *Root Mean Square Error*
RNA	Redes neurais artificiais
STF	Supremo Tribunal Federal
STJ	Superior Tribunal de Justiça
SVM	Máquina de vetor suporte
TRF	Tribunal Regional Federal
TRF4	Tribunal Regional Federal da 4ª Região
UF	Unidade da Federação

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta o contexto em que se situa a presente pesquisa, seguido da formulação do problema e os objetivos. Na sequência, é discutida a justificativa, a contribuição teórica proposta e são delineadas as limitações impostas à pesquisa. O capítulo se encerra com a estrutura da tese.

1.1 Contexto

Os princípios da administração pública, constantes na Constituição Federal, segundo Brasil (1988), são a legalidade, a impessoalidade, a moralidade e a publicidade. Em 1998, a emenda constitucional número 19 agregou a eficiência aos princípios existentes. A melhoria da eficiência está na agenda das organizações públicas jà que os recursos gastos são provenientes do pagamento de impostos.

Consoante à adição deste último princípio, a promulgação da Lei de Responsabilidade Fiscal (LRF), em 2000, estabeleceu um conjunto de normas e princípios, tais como os limites de gasto com pessoal, limites com o endividamento público e mecanismos de compensação para despesas de caráter permanente, limitando os recursos financeiros destinados à gestão pública. Além disso, a aprovação da Proposta de Emenda Constitucional (PEC) 241 limita o aumento do orçamento à inflação acumulada conforme o Índice Nacional de Preços ao Consumidor Amplo (IPCA), a partir de 2018. Assim, os órgãos públicos contam com recursos limitados para atender uma demanda irrestrita.

No contexto do Brasil, a busca por eficiência na gestão pública estende-se aos serviços judiciários, exemplificando o desbalanceamento entre demanda e capacidade no serviço público: em 2016, o índice de atendimento à demanda foi de 100,3%, de acordo com CNJ (2017) e, embora superior a 100%, não é suficiente para baixar um estoque de 79.662.896 processos pendentes no país.

Neste sentido, os três principais problemas do Poder Judiciário brasileiro, apontados pelo Ministério da Justiça, conforme CNJ (2014), são: a morosidade, o alto número de processos em estoque (ações judiciais ainda não finalizadas) e a falta de acesso à Justiça.

Especificamente no Brasil, segundo Jr (2012b), o Judiciário é considerado lento, ineficaz e caro, sendo a morosidade o principal aspecto da chamada Crise do Judiciário, pois diminui a sua eficácia enquanto mecanismo fomentador de cooperação e de desenvolvimento. Esta crise, de acordo com Sadek (2004), diz respeito a uma estrutura pesada, sem agilidade e incapaz de fornecer soluções em tempo razoável, previsíveis e com custos acessíveis.

O problema da morosidade não está presente apenas no judiciário brasileiro. O fun-

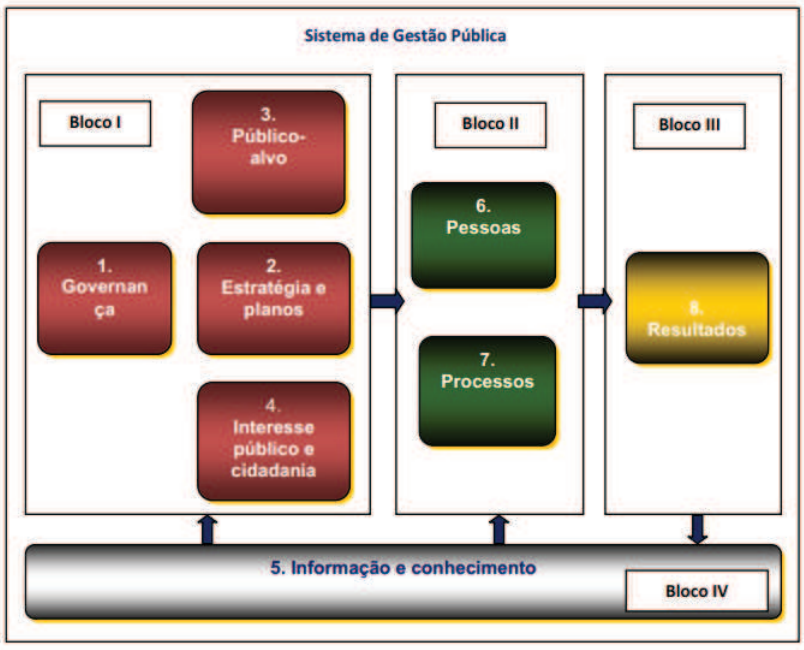
cionamento dos sistemas judiciais, de acordo com Martins (2009), é percebido como lento e inconsistente em sua forma de aplicação das leis, motivando a inclusão de reformas da Justiça na agenda do desenvolvimento econômico e social de instituições como o Banco Mundial e o Conselho da Europa. Como exemplo, pode ser citado o CEPEJ - *The European Comission for the Efficiency of Justice*, sendo que uma de suas prioridades é combater a morosidade nos sistemas judiciais europeus, com o estabelecimento de um grupo de trabalho para a gestão do tempo processual, chamado de centro Saturn. (FABRI; CARBONI, 2013).

Neste contexto, ao buscar eficiência, os órgãos públicos, como os Tribunais, podem contar com programas de gestão pública com foco em práticas de gestão. Em 2005, a partir do Decreto 5378, o Governo Federal lançou o Programa Nacional de Gestão Pública e Desburocratização - Gespública. O Gespública, segundo Ferreira (2009), está inserido em um contexto de mudança de paradigma administrativo: da administração burocrática para a administração gerencial, orientada por resultados.

O Gespública tem como base conceitual e metodológica um modelo de gestão chamado Modelo de Excelência em Gestão Pública (MEGP), que incorpora tanto a gestão técnica como a dimensão social, de acordo com Gespública (2009). As características principais do modelo são o foco em resultados para o cidadão e ser essencialmente público.

O modelo é composto por oito dimensões, conforme a Figura 1: Governança, Estratégia e Planos, Público Alvo, Interesse Público e Cidadania, Informação e Conhecimento, Pessoas, Processos, e Resultados.

Figura 1 – Sistema de Gestão Pública



Fonte: Gespública (2014)

Especificamente em relação à dimensão Informação e Conhecimento, segundo Ges-

pública (2014), esta representa a implementação de processos gerenciais que contribuam diretamente para a seleção, coleta, armazenamento, utilização, atualização e disponibilização sistemática de informações atualizadas, precisas e seguras, com o apoio da tecnologia da informação.

O Gespública prevê ainda que cada órgão ou entidade tenha uma Carta de Serviços. Em síntese, conforme Gespública (2014), a Carta de Serviços é o documento no qual o órgão ou a entidade pública estabelece o compromisso de observar padrões de qualidade, eficiência e eficácia na execução de suas atividades perante o seu público alvo e a sociedade.

Entre as finalidades da Carta de Serviços estão a avaliação contínua da gestão e o monitoramento do desempenho institucional mediante a utilização de indicadores, relativos tanto a processos de trabalho como a resultados.

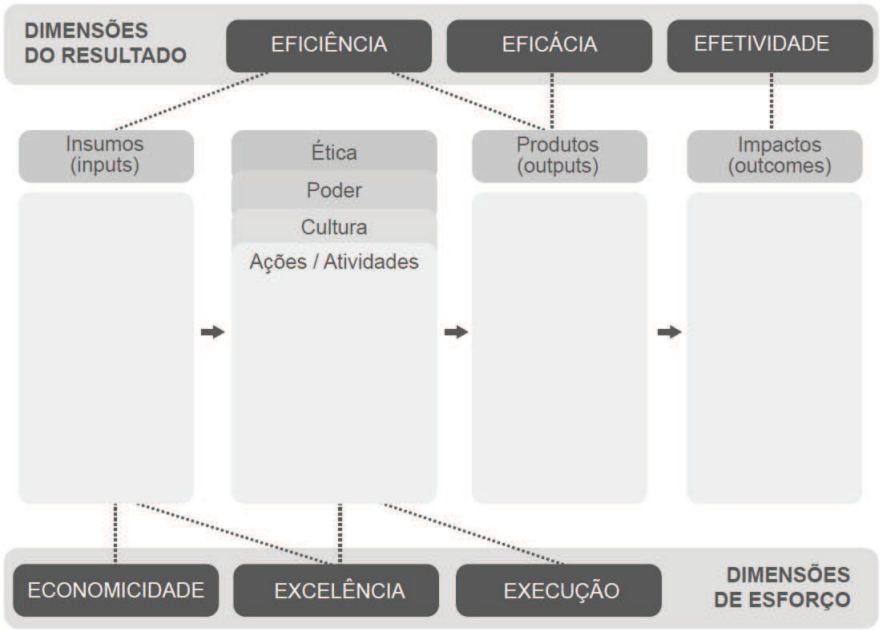
Os indicadores possuem duas funções básicas, de acordo com Gespública (2014): a de descrever o estado real dos acontecimentos e seu comportamento e a de analisar as informações presentes com base nas anteriores, de forma a realizar proposições valorativas. Assim, os indicadores permitem acompanhar o alcance das metas, identificar avanços, melhorias de qualidade, correção de problemas, necessidades de mudança, etc.

Conforme Palvarini (2010), a Secretaria de Gestão do Ministério do Planejamento, Orçamento e Gestão elaborou o Guia Referencial para Medição de Desempenho na Administração Pública como suporte à definição de indicadores. Este guia é um referencial metodológico que permite definir e mensurar desempenho, permitindo sua pactuação, avaliação e divulgação.

Para tanto, foi utilizado como metamodelo uma concepção de cadeia de valor que representa a atuação da ação pública desde a obtenção dos recursos até a geração dos impactos provenientes dos produtos/serviços e identifica seis dimensões do desempenho, chamadas de 6Es, agrupadas em dimensões de esforço (economicidade, execução e excelência) e dimensões de resultado (eficiência, eficácia e efetividade).

Este metamodelo permite que as organizações identifiquem distintos objetos de mensuração em cada dimensão, orientando a modelagem de indicadores e permitindo a construção de painéis de acompanhamento da gestão. O metamodelo preconiza, assim, regras básicas para a construção de modelos específicos de mensuração de desempenho. O modelo está mostrado na Figura 2.

Figura 2 – Cadeia de valor e os 6Es do Desempenho



Fonte: Martins e Marini (2010)

Já a Carta de Serviços, traz padrões de desempenho institucional na realização da atividade ou prestação de serviços mínimos:

- Tempo de espera para atendimento;
- Tempo de atendimento ou prazo máximo para a prestação do serviço/atividade;
- Prioridades de atendimento;
- Tratamento a ser dispensado aos usuários no atendimento;
- Condições mínimas a serem observadas pelas unidades de atendimento - em especial no que se refere à acessibilidade, limpeza e conforto;
- Requisitos básicos para o sistema de sinalização visual das unidades de atendimento;
- Informações sobre os meios de comunicação com os usuários, procedimentos para receber, atender, gerir e responder às sugestões e reclamações;
- Atividades (inclusive estimativas de prazos), mecanismos de consulta, por parte dos usuários, acerca das etapas, cumpridas e pendentes, para a realização do serviço/atividade solicitado.

Segundo Gespública (2014), entre os padrões de desempenho institucional constantes na Carta de Serviços, há destaque para o tempo de atendimento - prazo máximo para a prestação do serviço ou atividade, o qual pode ser considerado como um tempo de atravessamento máximo. Isto é, um indicador referente ao tempo para medir o desempenho institucional.

Para que indicadores, entre eles os de tempo de atravessamento, possam efetivamente fornecer informações para melhorar os resultados, eles necessitam ser confiáveis, estar disponíveis, atualizados, precisos e objetivos, servindo também para auxiliar na identificação de restrições e desbalanços de capacidade e demanda. Conforme Gespública (2014), as informações e dados devem ser utilizadas para subsidiar o planejamento, a avaliação, a tomada de decisões e a implementação de melhorias à medida em que elas definem tendências, projeções, causas e efeitos.

As tendências e projeções podem ser obtidas a partir de modelos, que são uma simplificação da realidade. Desta forma, a informação sobre tempo de atravessamento pode estar na forma de um modelo de sua predição, considerando o ajuste por meio de um montante considerável de dados disponíveis, mas brutos, que necessitam ser tratados a fim de gerar informação.

O Gespública enfatiza que os indicadores sejam relativos ao desempenho global institucional, aos seus processos internos, especialmente aos associados às atividades-fim da organização ou diretamente envolvidos no atendimento às necessidades dos seus cidadãos; aos públicos-alvos; aos servidores e ao ambiente externo, especialmente aos referenciais comparativos.

Desta forma, considerando os tempos de atravessamento como uma informação sobre a atividade-fim, diretamente envolvida às necessidades do cidadão, este indicador adequa-se ao programa Gespública. Um modelo de previsão deste indicador torna-se uma informação, um recurso passível de valoração e quantificação, algo que forneça subsídios para a tomada de decisões.

No contexto específico do judiciário brasileiro, de um sistema sobrecarregado, o conhecimento e a divulgação do tempo de atravessamento processual pode ser útil para a gestão judiciária à medida que permite o acompanhamento do desempenho da organização, auxilia no estabelecimento de padrões de qualidade e na identificação de problemas no fluxo da produção.

Considerando o tempo de atravessamento como o tempo da solicitação até a efetiva entrega, e portanto, identificado na cadeia de valor dos 6Es, ele pode auxiliar no planejamento dos seus recursos (influenciando a dimensão eficiência); e, também, pode auxiliar na tomada de decisão das partes em conciliar ou iniciar um litígio (atuando na dimensão efetividade). Reforça-se a importância da tomada de decisões a partir do conhecimento do valor deste indicador de desempenho.

Entre os indicadores que poderiam ser usados para avaliar a gestão judiciária, dentro da perspectiva dos 6 Es, o enfoque é dado ao tempo de atravessamento por que pode ser usado como fonte de informação para enfrentar o problema da morosidade, à medida em que pode ser utilizado para estabelecer parâmetros e acompanhar o desempenho da instituição.

Especificamente em relação à mensuração do tempo de processos judiciais, há vários termos e indicadores diferentes para a mensuração, assim como indicadores formais em diversos sistemas. Apresenta-se aqui um breve apanhado de indicadores relativos a tempos processuais.

1.2 Indicadores relativos a tempos processuais

São identificados os conceitos utilizados como indicadores de tempo processual da literatura relacionados à esfera judicial e, então, discutidos quanto à determinação de parâmetros de acompanhamento. Os conceitos de *disposition time*, *length of court proceedings* e *overall case length* são descritos a seguir.

- *Length of court proceedings*, em tradução literal é a duração dos processos judiciais.
- *Overall case length* ou tempo total de tramitação, é o tempo entre o registro inicial do processo até o seu final. De acordo com Dalton e Singer (2014), esta é uma medida importante da eficiência da corte, embora esta variável sozinha não capture adequadamente a informação sobre satisfação com o sistema judicial, mas é um componente da satisfação, e considerada muito importante. No Brasil, este indicador é conhecido como tempo de tramitação dos processos baixados.
- *Disposition time*. Há dois conceitos para esta medida: é considerado como o número de dias entre a data de acusação e a data da decisão final, segundo Walsh et al. (2015), tal como a data da negociação ou julgamento. Porém, este indicador também é definido como o número de dias necessários para que um processo pendente seja resolvido, (LEPORE; METALLO; AGRIFOGLIO, 2012). Este segundo conceito não leva em conta a duração do processo e é calculado dividindo-se 365 pelo indicador *case turnover ratio* (a razão de casos resolvidos e o número de casos não resolvidos).
- Em processos criminais, o tempo em que um caso permanece no sistema judicial (*length court proceedings*) é composto pelo tempo atual de processamento, isto é, o tempo que um juiz e outros participantes da corte, como servidores, destinam para um caso particular – e seus tempos de espera, segundo Shamir e Shamir (2012), mais parecido com o conceito de tempo de atravessamento oriundo da engenharia de produção.

Destes conceitos, *length of court proceeding* e *overall case length* são medidas de tempo de atravessamento, enquanto o *disposition time* pode ser compreendido como uma medida calculada para estabelecer prazos de tempo processual. Para acompanhar esta medida, há um outro indicador, a porcentagem de casos com tempo maior do que prazos estabelecidos ou *case backlog*.

No Brasil, o CNJ (Conselho Nacional de Justiça) formulou, por meio da Resolução 76/2009, republicada em 2015, segundo CNJ (2009), uma série de indicadores para acompanhar o desempenho do judiciário brasileiro. Os indicadores, entre outras funções, são utilizados na tomada de decisão sobre a necessidade de aprimoramento da estrutura dos tribunais e abrangem as dimensões litigiosidade, recursos humanos, finanças, e tempo de tramitação processual. Estes indicadores são divulgados pelo CNJ, relativos ao ano anterior, ou seja, retrospectivos. Os de litigiosidade e de tempos de tramitação são classificados em criminais e não criminais.

Especificamente na dimensão tempo de tramitação processual, de acordo com CNJ (2009), há a publicação dos seguintes indicadores, por meio das medidas média, mediana e desvio padrão:

- Tempo de tramitação dos processos baixados;
- Tempo de tramitação dos processos pendentes;
- Tempo da decisão (terminativa, sentença);
- Tempo da suspensão (total, em função de repercussão geral, de recurso repetitivo).

A primeira medida é o correspondente ao tempo de atravessamento. A segunda medida é o prazo de estocagem, ou seja, o tempo de estoque no tempo t, o tempo da decisão refere-se ao tempo do órgão julgador, incluindo o tempo do magistrado, e por fim, o tempo da suspensão refere-se ao tempo de espera no qual o processo não está em movimento e nem acabado.

Após a descrição dos indicadores sobre tempo de atravessamento processual para a Gestão Judiciária, são abordados, na próxima seção, os trabalhos acadêmicos realizados com o foco em tempo de atravessamento.

1.3 Pesquisas sobre tempo de atravessamento processual

Nesta seção apresentam-se alguns dos trabalhos acadêmicos já realizados sobre indicadores de tempo de atravessamento processual e seus modelos de predição, principalmente no Brasil.

Em relação à pesquisa acadêmica no Brasil, há tanto estudos com enfoque em processos de trabalho como em modelos de previsão e avaliação. Estes trabalhos estão descritos a seguir. Na temática da eficiência, destacam-se os trabalhos de:

- Bordasch (2009), Neto (2010) e Clemes (2010), que estudaram o tempo de atravessamento do ponto de vista dos processos, com sugestões de redesenho;
- Guimarães (2014), a partir do tempo de atravessamento, enfatizou a eficiência como combate à morosidade;
- Nogueira (2010) abordou o estabelecimento de metas e indicadores para o judiciário relacionados a tempo;
- Olivieri (2010) e Baldan (2011) estudaram a contribuição do processo eletrônico para a redução do tempo de atravessamento processual.

Quanto a modelos de predição de tempo de atravessamento, há trabalhos tanto no Brasil como no exterior, usando principalmente análise de regressão e mineração de dados, destacando-

se as redes neurais para sua construção. Estes trabalhos podem fazer parte do tema jurimetria, uma disciplina emergente.

Jurimetria, conforme Loevinger (1948), é definida como a análise quantitativa do comportamento judicial, a aplicação da teoria da informação e comunicação à expressão legal, o uso da lógica matemática na lei, a recuperação de dados legais por meios eletrônicos e mecânicos e a formulação de um cálculo da previsibilidade jurídica. Couto e Oliveira (2016) citam a jurimetria como uma importante ferramenta auxiliar na análise da necessidade de compatibilidade entre o fluxo, a gestão de processos e a política judicial (planejamento estratégico, organização e gestão do judiciário). Podem ser destacados os seguintes trabalhos:

- Zhou (2008) propôs um modelo para predizer a duração do processo baseado nos dados do Texas Departamento de Seguros de Responsabilidade Civil para o período de 1988-2005, por meio da análise de regressão de Cox. Quanto às variáveis respostas utilizadas, o tempo processual foi dividido em dois: tempo antes do julgamento e depois do julgamento. O autor dividiu o tempo total em estágios do processo e, então, usou como variável resposta o tempo discreto mensurado em meses. O autor mostrou os tempos médios e medianos dos estágios do processo e o tempo total. Em todos os estágios, o tempo médio foi superior ao mediano, porém, superior ao desvio padrão. Não é apresentada uma medida para o erro de ajuste.
- Dalton e Singer (2014) utilizaram análise de regressão para predizer o tempo total de processos civis. Os autores utilizaram dados do IAALS (*Advancement of the American Legal Systemat the University of Denver*), através de um banco de dados referente a casos encerrados no período de 1 de outubro de 2005 a 30 de setembro de 2006, com exceções, com aproximadamente 6700 casos em 7 distritos federais. Os autores escolheram a duração total do caso como variável dependente porque essa era uma medida disponível em todos os casos. O objetivo foi prever a duração de um caso com base em uma função do número de advogados e do tamanho dos tribunais. Para tanto, os autores testaram quatro modelos de regressão. O modelo com melhor ajuste foi obtido através da comparação da medida deviance, a qual resultou em 95,988 (o tempo médio foi de 325 dias).
- Spurr (2000) analisou os fatores que determinam a duração de um litígio com o foco em dois assuntos particulares: negligência médica e ferimentos pessoais, usando como técnica de análise a regressão linear. O autor utilizou como variável o tempo para liquidação, utilizando uma transformação logarítmica. A técnica usada foi dos mínimos quadrados ordinários. Como R^2, o autor chegou ao valor de 0,1192 (o logaritmo do tempo médio é de 0,814).
- Bielen, Marneffe e Vereeck (2015) utilizaram mínimos quadrados ordinários, com variáveis baseadas na literatura, para categorizar os determinantes do tempo processual, sendo o primeiro artigo a analisar o tempo de processos civis na 1^a Instância Belga.

- Shamir e Shamir (2012) desenvolveram um modelo para estudar os fatores que afetam o tempo de atravessamento processual e como este atraso afeta o bem estar social. Os autores utilizaram a teoria das filas para representar a forma como a congestão nas cortes é criada, sendo possível, através do modelo, que os tomadores de decisão efetivamente comparem diferentes propostas políticas e examinem seus efeitos no tempo processual.
- Schneider (2003) aplicou mineração de dados no Tribunal Estadual do Rio Grande do Sul. Os dados foram divididos em dois grupos. O primeiro, referiu-se à incidência de processos de acordo com a classificação processual de seção, classe e especialização e à relação quanto a comarca de origem; tempo de tramitação processual (tempo de atravessamento); tipo de sentença; e influência de ocorrência de audiência. O segundo grupo tratou o perfil dos réus em processos criminais considerando sexo, grau de instrução, estado civil, cor, profissão e natureza do processo.
- A partir de dados de processos do Fórum Trabalhista de São José dos Pinhais, Pavanelli e Pavanelli publicaram uma série de trabalhos. Inicialmente, Pavanelli (2007) estudou o tempo de tramitação processual. O tempo de tramitação processual foi usado como variável resposta, transformada através da divisão do valor pelo tempo máximo. O autor comparou diversas arquiteturas de redes neurais para modelar um tempo com valores originais no intervalo entre 2 e 94 meses.
- Pavanelli et al. (2011) aplicaram redes neurais artificiais (RNAs) e regressão linear múltipla com os objetivos de rever o tempo de duração das seções de audiências para gerar uma agenda inteligente a fim de melhorar o agendamento dos horários dos magistrados.
- Pavanelli, Pavanelli e Costa (2013) aplicaram as técnicas de redes neurais artificiais, regressão linear múltipla e árvores de decisão como técnicas alternativas para prever o tempo de duração do trâmite de processos. Os autores concluíram que as diferentes técnicas apresentaram resultados satisfatórios, mas as RNAs apresentaram um desempenho superior.
- Silva, Cunha e Talon (2013) utilizaram mineração de dados para analisar os processos jurídicos do Tribunal de Justiça de São Paulo utilizando o *business intelligence* disponível no SQL Server 2008, com o propósito de estudar o tempo de tramitação processual. Os autores verificaram qual assunto tem maior probabilidade de processos com longa duração e as comarcas.
- Rosa et al. (2017) estudou os fatores associados à variância no tempo de tramitação necessário ao fim do litígio dos processos dos Juizados Especiais Cíveis do Poder Judiciário de Santa Catarina. Os autores agruparam os processos em níveis conforme o assunto e a comarca usando o Modelo Linear Hierárquico com Efeitos Cruzados para avaliar o impacto de cada variável no tempo de tramitação.
- Gruginskie e Vaccaro (2018) propuseram um modelo para predição da faixa de tempo de atravessamento na Justiça Federal brasileira usando, como variável resposta, o tempo

estratificado em faixas.

Destaca-se que alguns dos modelos de previsão gerados (Schneider (2003); Pavanelli (2007), Pavanelli (2008); Silva, Cunha e Talon (2013)), serviram como fonte de informação para as organizações estudadas identificarem problemas. Apesar destes trabalhos terem proposto modelos para a análise do tempo de atravessamento processual, pode-se discutir as variáveis usadas, suas relações, além da comparação com o desempenho no uso de outras técnicas de análise. Os trabalhos consultados utilizaram-se, principalmente de análise de regressão e redes neurais em estudos retrospectivos. Outras técnicas também podem ser usadas, entre elas, a Máquina de Vetor Suporte (SVM) e a Análise de Sobrevivência, embora esta última tenha sido usada por (ZHOU, 2008).

1.4 Problema de Pesquisa

Como foi visto nas seções anteriores, de acordo com CNJ (2014) e Fabri e Carboni (2013), tanto no Brasil como na Europa, a morosidade é apontada como um dos principais problemas do judiciário. Entretanto, já existe uma preocupação com a duração de processos, através do estabelecimento de prazos razoáveis. No Brasil, esta definição é colocada pela proposição da meta 2 do CNJ, de julgar os processos mais antigos.

Para acompanhar, medir e estabelecer medidas que definam a morosidade, torna-se fundamental, tanto para a gestão judiciária como para o jurisdicionado, o acompanhamento do tempo de atravessamento. Para tanto, é fundamental a discussão dos indicadores que o medem, os modelos usados para a sua previsão e o desenvolvimento de novos modelos.

Na literatura, ainda incipiente sobre tempo de atravessamento processual, existem indicadores para acompanhamento, entre eles o tempo de duração do processo e o *disposition time*. Entretanto, no Brasil, percebe-se interesse em estudar a modelagem de tempo de tramitação de processos utilizando, principalmente, redes neurais.

Neste trabalho são abordados apenas processos civis, em vez da abordagem de processos civis e criminais, por dois motivos: i) os processos civis seguem o Código de Processo Civil e o [1]fluxo, assim como as decisões, são diferentes de processos criminais; ii) tanto nos trabalhos como nos indicadores pesquisados, há distinção entre os processos criminais e não criminais, reforçando esta diferença.

Os processos não criminais referem-se ao meio ambiente, previdência social, direito tributário, licitações, contratos de financiamento habitacional firmados com empresas públicas ou autarquias, questões relativas a concursos e a imóveis da União, entre outras. De forma geral, os processos civis abragem todos os ramos de direito não criminal: previdenciário, ambiental, administrativo, tributário, marítimo, etc. Em matéria penal, a Justiça Federal tem na sua competência

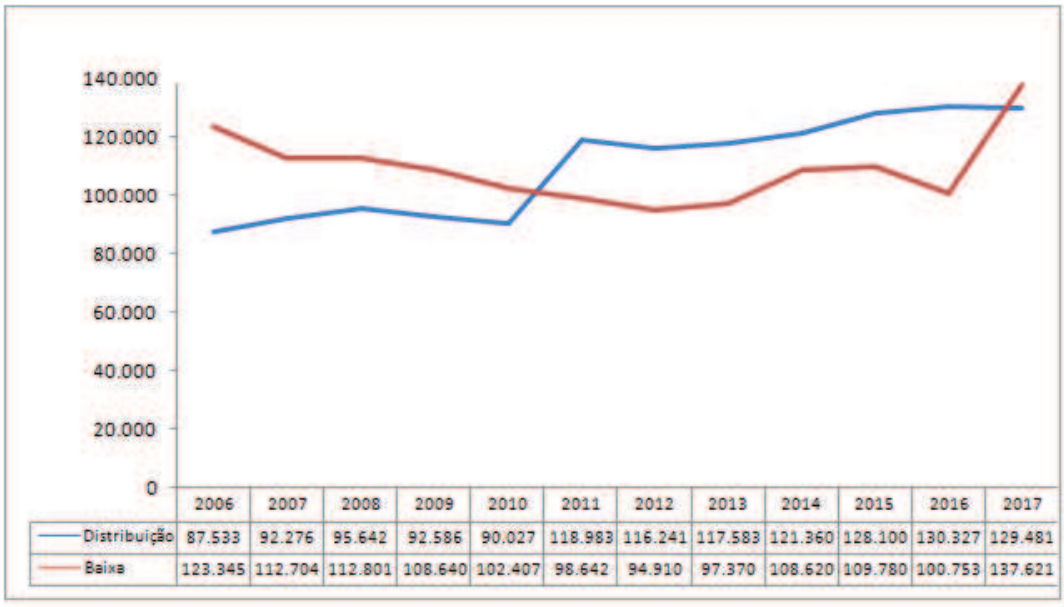
[1]

o julgamento de crimes fiscais, de lavagem de dinheiro, de tráfico internacional de entorpecentes entre outros. São comuns as ações de massa, como as que versam sobre a correção monetária do FGTS, as ações previdenciárias, os processos tributários e os que tratam dos financiamentos da casa própria.

Em relação ao acesso aos dados, foi permitida a consulta das variáveis constantes no *datawarehouse* institucional do Tribunal Regional Federal da 4ª Região (TRF4), sendo este o órgão de estudo. O TRF4 é um órgão de segunda instância, responsável por 28% da demanda da justiça federal não criminal de segundo grau em 2016 (CNJ, 2016b). Ao TRF4 compete julgar recursos e processos originários, embora de acordo com CNJ (2016b), 98% dos casos novos não criminais em 2016 foram recursais.

Os problemas do judiciário podem ser comprovados no TRF4: demanda maior que a capacidade; recursos limitados; morosidade e alto número de processos em estoque. Em relação ao desbalanceamento entre demanda e oferta, conforme a Figura 3, a quantidade de processos distribuídos (demanda) foi superior a 100.000 desde 2011. Porém, a baixa não está acompanhando: de 2010 a 2016 a demanda, representada pela linha azul, é maior que a baixa de processos, representada pela linha vermelha. Embora a demanda tenha se apresentado crescente desde 2006, os recursos, como a quantidade de Magistrados, continua a mesma desde então.

Figura 3 – Processos distribuídos e baixados no TRF4 de 2006 a 2017 - Competência não criminal



	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Distribuição	87.533	92.276	95.642	92.586	90.027	118.983	116.241	117.583	121.360	128.100	130.327	129.481
Baixa	123.345	112.704	112.801	108.640	102.407	98.642	94.910	97.370	108.620	109.780	100.753	137.621

Fonte: Dados fornecidos pelo TRF4

A demanda maior que a capacidade faz aumentar a quantidade de processos não criminais em tramitação, ou estoque processual. Em 2016, o estoque era de 110.016, superior a quantidade de processos baixados naquele ano. Quanto à morosidade, o tempo médio entre autuação e baixa dos processos não criminais foi de 791 dias em 2017, o equivalente a 26 meses.

Desta forma, dado a preocupação com a morosidade, a preocupação de órgãos nacionais com a celeridade, a publicação de trabalhos sobre tempo de atravessamento brasileiros, e a permissão para a realização de pesquisa em um órgão do judiciário federal, é proposta a seguinte questão norteadora de pesquisa: Como fazer a previsão do tempo de atravessamento processual, de ações cíveis na justiça federal, considerando as características dos processos no momento de autuação, a partir de variáveis disponíveis em bancos de dados?

Face a essa breve apresentação do problema sob análise, o restante do capítulo apresenta os objetivos, a justificativa da pesquisa, as delimitações impostas para sua realização e a contribuição téorica dela advinda.

1.5 Objetivos

Para responder à questão de pesquisa, o objetivo geral proposto é: Estruturar modelos de previsão de tempo de atravessamento processual de ações judiciais civis na justiça federal brasileira, utilizando informações históricas e considerando as características do processo no momento de autuação, para comparação dos modelos e que sirva de informação para as partes processuais e administração.

Os objetivos específicos são:

- Construir quatro modelos preditivos usando, em pelo menos dois deles, uma variável quantitativa como resposta e, em pelo menos um, uma variável qualitativa estratificada em faixas de tempo;
- Testar os modelos no contexto da Justiça Federal brasileira de 2^o grau da 4^a região;
- Discutir os modelos obtidos quanto i) ao tipo de estudo (retrospectivo, prospectivo), ii) à variável resposta e iii) às covariáveis usadas;
- Analisar o tempo de atravessamento processual em relação às covariáveis analisadas;
- Verificar a possibilidade da proposta da duração razoável para tempos de atravessamento.[2]

1.5.1 Abordagem

Objetivou-se construir quatro modelos preditivos pois o problema apresentado pode ser abordado por meio de diferentes técnicas, entre elas, mineração de dados, análise de sobrevivência, método Delphi, análise de regressão.

Como o tempo está medido em dias, em uma escala quantitativa discreta, a opção foi ajustar um maior número de modelos comparando o tempo de atravessamento na escala original ou transformada. A análise por meio da variável categorizada em faixas de tempo é uma alternativa e por isso, foi analisado apenas um modelo nesta escala.

[2]

Assim, são propostos, os seguintes modelos para o tempo em escala quantitativa, transformada ou original: i) análise de sobrevivência, ii) análise de regressão com o uso de redes neurais, iii) análise de regressão com o uso de máquina de vetor suporte e, para a variável categorizada em faixas de tempo, iv) a classificação com o uso de máquina de vetor suporte.

No caso de emprego da técnica de análise de sobrevivência, a variável resposta é o tempo, tornando esta técnica a forma usual de analisar tempos. Como a análise de sobrevivência é utilizada em vários campos de estudo, segundo Allison (2010), recebe diferentes nomes como, por exemplo, análise de eventos históricos (sociologia), análise de confiabilidade (engenharia), análise de duração (economia). De acordo com Liu (2012), a característica de mudança de estado faz com que análise de sobrevivência seja similar a técnicas estatísticas com resultado qualitativo, tal como regressão logística ou probito.

Em relação ao plano de observação para análise de sobrevivência, segundo Allison (2010), o melhor é o prospectivo, sendo que para algumas situações, pode ser retrospectivo.

Outra técnica para a proposta de modelos é a Máquina de Vetor Suporte (SVM), tanto para regressão como para classificação. De acordo com Han, Kamber e Pei (2011), SVM tem sido aplicada a diversas áreas, como reconhecimento de dígitos manuscritos, reconhecimento de objetos e identificação de palestrantes, e conforme Hornik, Meyer e Karatzoglou (2006), em testes de previsão de séries temporais de referência, biotecnologia e astrofísica. Como desvantagens, Han, Kamber e Pei (2011) apontam o tempo de treinamento lento, e como vantagens, a alta precisão devido à sua capacidade de modelar limites complexos de decisão não-lineares; e uma menor propensão a sobre-ajustamento.

Como foi citado, redes neurais já foram utilizadas para previsão de tempo de atravessamento processual. De acordo com Hastie, Tibshirani e Friedman (2009), redes neurais são especialmente eficazes em problemas com uma alta relação sinal-ruído e em configurações em que a previsão sem interpretação é o objetivo, sendo menos eficazes para problemas em que o objetivo é descrever o processo físico que gerou os dados e os papéis das entradas individuais. Por outro lado, segundo Han, Kamber e Pei (2011), redes neurais envolvem longo tempo de treinamento.

Conforme Han, Kamber e Pei (2011), as redes neurais têm sido aplicadas com êxito em áreas tais como reconhecimento de caracteres manuscritos, patologia e medicina laboratorial, e também para treinamento de máquina para pronunciar um texto em inglês.

As técnicas, embora possam ser usadas para o mesmo fim, apresentam enfoques diferentes: análise de sobrevivência, segundo Liu (2012), refere-se a fazer comparações entre grupos/categorias de uma população ou examinar as variáveis que influenciam o processo de sobrevivência, enquanto redes neurais, segundo Han, Kamber e Pei (2011), são criticadas por sua pobre interpretabilidade em relação às variáveis. Por outro lado, modelos de análise de sobrevivência conseguem adaptar mudanças contemporâneas ao modelo, uma vez que tem um

enfoque prospectivo.

1.6 Justificativa e Contribuição

Esta seção destaca a importância das instituições públicas e do Poder Judiciário para a sociedade e economia; a importância do indicador do tempo de atravessamento para a aferição do desempenho do Poder Judiciário; a relevância da tese para a sociedade, Poder Judiciário, instituição e engenharia de produção; e por último, as áreas de engenharia de produção relacionadas com o trabalho de pesquisa.

Bresser-Pereira (2008) enfatiza a importância das instituições públicas e conclui que elas são fundamentais na promoção do desenvolvimento econômico, sendo o próprio Estado a instituição central das sociedades modernas à medida em que dá origem às instituições normativas formais (Leis). O sistema judicial, de acordo com Falavigna et al. (2015), é essencial para a proteção dos direitos formais dos cidadãos, na interação com os Poderes Legislativo e Executivo, à medida que contribui para obter resultados sociais, na promoção do desenvolvimento sustentável, na aplicação dos direitos de propriedade privada, na confiabilidade dos sistemas econômicos e no cumprimento dos contratos.

De forma geral, conforme Pereira (2010), a morosidade, os custos de acesso e a imprevisibilidade das decisões judiciais são empecilhos para o crescimento econômico. Segundo Dakolias (1996), a aplicação das normas e Leis de forma previsível e eficiente tem um reflexo positivo na economia de um país à medida que as novas relações comerciais demandam decisões imparciais e a resolução de seus conflitos através das instituições formais.

Desta forma, uma má atuação do Poder Judiciário tem reflexos na economia do país, pois desestimula as transações comerciais, adicionando-lhes riscos e custos, reduzindo o tamanho do mercado e sua competitividade.

Em contrapartida, de acordo com Dakolias (1996), se o judiciário possuir: a) previsibilidade nos resultados dos processos; b) acessibilidade; c) tempo razoável de julgamento; e d) recursos processuais adequados, haverá garantia de um ambiente institucional estável onde os resultados econômicos a longo prazo podem ser avaliados.

Já existe uma preocupação no País com estes problemas, principalmente com a morosidade. O artigo 5^{o} da Emenda Constitucional 45, de 2004, impõe a razoável duração do processo e os meios que garantam a celeridade de sua tramitação. (BRASIL, 2016). Outra preocupação foi a instituição, em 2009, pelo Conselho Nacional de Justiça, das metas de nivelamento do Poder Judiciário, com os objetivos de proporcionar agilidade e eficiência à tramitação processual, melhorar a qualidade do serviço jurisdicional prestado e ampliar o acesso do cidadão brasileiro à justiça. (CNJ, 2016a).

Para combater a morosidade, é necessário quantificá-la. No Brasil, a obtenção de indica-

dores de tempo de atravessamento bem como os modelos para predição têm apoio do Ministério de Planejamento e Gestão: de acordo com Gespública (2014), à medida que as informações geradas são obtidas de forma sistemática, alinhadas às necessidades da organização, a partir de sistemas estruturados e adequados, elas podem gerar projeções, relações de causa e efeito que serão subsídio para o planejamento, à avaliação e à tomada de decisão. O uso sistemático destas informações produzem conhecimento que fornecem à organização a capacidade para agir e inovar.

Reforçadas a importância da gestão judiciária e considerado o problema previamente apresentado, pode-se citar as contribuições deste estudo para a academia, para a gestão judiciária e outras perspectivas: da sociedade, das partes interessadas, dos advogados e procuradores, do sistema judicial e da engenharia de produção.

Da perspectiva da sociedade, disponibilizar a informação pode aumentar a confiança no judiciário à medida que o conhecimento pode influenciar na decisão de resolver um conflito civil. De acordo com Cunha et al. (2015), no primeiro semestre de 2017, o índice de confiança na Justiça, que varia de 0 a 10, foi igual a 4,5; menor que o índice de 2016, de 4,9. Conforme Cunha et al. (2015), a confiança em uma instituição significa identificar se o cidadão crê que a instituição cumpre a sua função com qualidade, se os custos são menores que os benefícios e se a instituição é levada em conta no dia-a-dia do cidadão comum. O índice é composto por dois subíndices: percepção (referente à confiança, à rapidez na solução dos conflitos, aos custos do acesso, à facilidade no acesso, à independência política, à honestidade e à capacidade para solucionar os conflitos) e comportamento (relativo a situações em que o cidadão procuraria o judiciário para a resolução de conflitos tal como o direito do consumidor).

A partir da informação às partes interessadas, o trabalho pode ser útil porque auxilia na tomada de decisão em conciliar, entrar com uma ação ou não, poupando seu tempo e seus custos. Entende-se por partes interessadas, o autor da ação, o que se sentiu lesado pelo descumprimento de um direito, em processo civil, ou aquele que procura a reparação de um direito; e o réu, aquele que negou um direito. Pode haver, em uma mesma ação, mais de um autor ou mais de um réu.

Os advogados e procuradores são o elo entre a parte e o sistema judicial, e podem informar melhor às partes sobre o tempo necessário com o objetivo de os auxiliar na tomada de decisão em conciliar ou litigar. O Poder Judiciário é o agente que representa o sistema judicial, ou seja, a instituição mediadora do conflito, representada por seus magistrados e servidores, ambos com número insuficiente, alta carga de trabalho e falta de treinamento. Neste sentido, compreender a economia do atraso judicial e melhorar o desempenho do tribunal são de importância vital para o sistema legal.

Do ponto de vista acadêmico, no final da década de 2000, Nogueira e Pacheco (2009) identificaram uma lacuna existente na literatura sobre gestão judiciária. Desde então, foram publicados diversos trabalhos conforme exposto na seção 1.3. Da mesma forma, Pavanelli et al. (2011) salientam não ter encontrado trabalhos de previsão relacionados a problemas da Justiça,

no caso, a Justiça do Trabalho, sendo estes autores, um dos pioneiros na publicação de trabalhos referentes a previsão de tempo na justiça brasileira.

A Engenharia de Produção, através de conhecimentos oriundos das áreas de Gestão de Operações e Pesquisa Operacional (mais especificamente Gestão de Informação e Mineração de Dados), poderá fornecer estimativas do tempo e contribuir para melhorar a eficiência do Poder Judiciário. Sob este ponto de vista, de acordo com a classificação de áreas da Engenharia de Produção, proposta pela ABEPRO, Associação Brasileira de Engenharia de Produção, este trabalho pode ser considerado dentro das seguintes áreas:

- área: Pesquisa Operacional, sub-áreas Processos Decisórios, Análise da demanda; Inteligência Computacional. Especificamente na área de inteligência artificial, considerando que Agrupamento, Classificação e Reconhecimento de Padrões são exemplos de aplicação de inteligência computacional;

- área: Engenharia Organizacional, sub-área Gestão da Informação. À medida que se trabalhará com os fluxos formais de informação e se usará os recursos de informação relevantes para o Poder Judiciário.

Os trabalhos consultados e apresentados anteriormente enquadram-se nas mesmas áreas de Engenharia de Produção deste trabalho. Sob o ponto de vista acadêmico, mesmo que já tenham sido publicados artigos e dissertações sobre modelos para estimar o tempo de processos judiciais, a discussão pode ser ampliada em relação ao tipo de estudo, às covariáveis e à variável resposta.

Em síntese, esta tese tem as seguintes pretensões:

- Realizar uma revisão sobre os trabalhos publicados sobre tempo de atravessamento processual, principalmente no Brasil, contribuindo, desta forma, para a literatura de tempos processuais e gestão judiciária. À medida que compara e estrutura modelos para a previsão do tempo de atravessamento processual de ações civis da justiça federal, complementa a lacuna dos estudos sobre mineração de dados no Poder Judiciário, abordando a Justiça Federal, e também na análise do tempo por faixas. A exceção do trabalho de Gruginskie e Vaccaro (2018), o qual pode ser considerado parte desta tese, ainda não há outras publicações. A intenção é aproximar-se aos trabalhos de Schneider (2003), Pavanelli (2008), Pavanelli (2007), Pavanelli et al. (2011), Pavanelli, Pavanelli e Costa (2013), Silva, Cunha e Talon (2013), à medida que pretende aplicar mineração de dados ao judiciário;
- Contribuir para a tomada de decisão das partes processuais e da gerência. As partes processuais, à medida que conhecem o tempo estimado para a duração processual, podem decidir entrar em conflito ou conciliar. A informação da previsão de tempo de atravessamento, para a administração, é útil para comparar os resultados com os parâmetros estabelecidos de duração processual, se existirem. Para a Gestão Judiciária, o conhecimento do tempo de atravessamento pode ser útil na aplicação da Lei de Little, conforme trabalho publicado

por Little e Graves (2008). Esta Lei relaciona três grandezas fundamentais: o trabalho em processo (WIP), o tempo de ciclo (CT) e a taxa de produção em um sistema, subsistema ou mesmo recurso. A partir da análise da Lei de Little, é possível analisar a viabilidade de estabelecimento de parâmetros para a duração de processos;

- Contribuir para um tema emergente, a jurimetria, cujo conceito tem sido aplicado no Brasil nos últimos anos.

1.7 Delimitação

A viabilidade operacional deste trabalho dependeu de delimitações. O trabalho foi baseado em dados existentes em sistemas, onde há o registro eletrônico dos processos existentes e seus diversos desfechos e roteiros, ao longo de sua vida. Da compreensão do processo judicial como fluxo de produção e da análise dos bancos de dados existentes, foram extraídos os modelos conceituais para geração da informação pretendida nesta tese. No entanto, acredita-se que uma limitação possa existir referente aos dados: a ausência de dados de variáveis relevantes nos sistemas de controle de armazenamento e que possam interferir no tempo processual representa perda de efetividade das estimativas pretendidas.

Considerando que tratam-se de modelos estocásticos empíricos, há limitações nas aplicações. À medida que estes modelos tentam representar um sistema com base em variáveis disponíveis, são uma representação simplificada do todo e, se variáveis importantes não são incluídas, o modelo acaba não sendo satisfatório. Como explicação, pode-se citar a opinião de Box sobre modelos: segundo Box e Jenkins (1976), embora todos os modelos estão errados, eles podem ser úteis. Assim, mesmo que as medidas indiquem uma baixa qualidade de ajuste, os modelos podem ser úteis para a compreensão do tempo de atravessamento processual.

1.8 Estrutura da Tese

A Tese está estruturada em seis capítulos, sendo este introdutório, seguido de dois capítulos de referenciais teóricos, um capítulo de método, um de resultados e uma conclusão.

O Capítulo 2 é o primeiro do referencial. Neste capítulo, são apresentadas a organização da justiça brasileira e da Justiça Federal para explicar as características do judiciário no Brasil. Da mesma forma, são apresentados os principais problemas do judiciário e suas relações para dimensionar os problemas enfrentados e reforçar a importância de lidar com a mensuração do tempo de atravessamento. Neste mesmo capítulo, ainda são abordados os indicadores de desempenho, e são analisados trabalhos identificados na mesma temática desta tese e seus resultados.

O Capítulo 3, também de referencial teórico, apresenta conceitos de mineração e ciência de dados. Ele constitui a base teórica das técnicas que se pretende utilizar para a modelagem

referente às técnicas usadas nos capítulos posteriores.

O Capítulo 4 traz o método de pesquisa usado, bem como a sua justificativa. No capítulo também é apresentado o método de trabalho, com a descrição de todos os passos necessários para alcançar os objetivos propostos no Capítulo 1.

O Capítulo 5 apresenta os resultados da análise, as conclusões e perspectivas de trabalhos futuros. O último capítulo traz a conclusão e encerramento. Por fim, são apresentados os elementos pós-textuais: referências e apêndices.

2 ORGANIZAÇÃO DA JUSTIÇA BRASILEIRA

Este capítulo apresenta, na primeira seção, o referencial relativo à Estrutura do Poder Judiciário, com o objetivo de compreender a organização da Justiça Brasileira. A seguir, há uma breve descrição da Justiça Federal e da 4^a região, para contextualizar o âmbito estudado. Após, são comentados os principais problemas do judiciário brasileiro e referidos os indicadores judiciais utilizados no Brasil e na Europa. Por último, é apresentada uma síntese dos trabalhos apresentados sobre o tempo de atravessamento no judiciário.

2.1 Estrutura do Judiciário Brasileiro

O Poder Judiciário brasileiro tem como função, segundo o artigo 2 da Constituição Federal (Brasil (1988)), a administração da Justiça e o cumprimento da Constituição, sendo independente dos demais Poderes. De acordo com o artigo 92 da Constituição Federal, o Poder Judiciário é composto: a) pelo Supremo Tribunal Federal (STF); b) pelo Conselho Nacional de Justiça (CNJ); c) pelo Superior Tribunal de Justiça (STJ); d) pela justiça comum, composta por d.1) Tribunais Regionais Federais e Juízes Federais; d.2) Tribunais e Juízes dos Estados e do Distrito Federal; e e) pela justiça especializada: e.1) Tribunais e Juízes do Trabalho; e.2) Tribunais e Juízes Eleitorais; e.3) Tribunais e Juízes militares. Em resumo, a organização judiciária brasileira é composta pela justiça comum e pela justiça especializada, pelo STF, CNJ e STJ. O STF é a mais alta instância do Poder Judiciário do Brasil e acumula competências de Suprema Corte (tribunal de última instância) e tribunal constitucional (que julga questões de constitucionalidade independentemente de litígios concretos). Abaixo do STF está o STJ, responsável por fazer uma interpretação uniforme da legislação federal.

A justiça especializada é formada pelos Tribunais Eleitorais, do Trabalho e Militares, com as competências respectivas. A justiça comum é composta pela Federal e a Estadual. De acordo com CNJ (2010), a Justiça Federal é responsável por processar e julgar as causas em que a União, suas entidades autárquicas e empresas públicas federais figurem como interessadas na condição de autoras ou rés, além de outras questões de interesse da Federação, enquanto à Justiça Estadual compete julgar as matérias que não são de competência da Justiça Federal ou de qualquer outra justiça especializada. Geralmente, os processos se originam na primeira instância, podendo ser levados, por meio de recursos, para a segunda instância. Conforme Leiria (2006), os processos que ainda vão ser objeto de recurso, voltam aos Tribunais Superiores (STF, STJ ou Tribunais Especializados) em terceira ou quarta instância. A Figura 4 mostra a organização da justiça brasileira.

Figura 4 – Organização da Justiça Brasileira

Fonte: CNJ (2010)

De acordo com Brasil (2016), o Conselho Nacional de Justiça (CNJ) foi criado pela Emenda Constitucional 45, com o objetivo de melhorar o trabalho do sistema judiciário brasileiro, no tocante ao controle e à transparência administrativa e processual. Entre as atribuições do CNJ estão, conforme CNJ (2016a), definir o planejamento estratégico, os planos de metas e os programas de avaliação institucional do Poder Judiciário; elaborar e publicar relatório estatístico sobre a movimentação processual e outros indicadores sobre a atividade jurisdicional brasileira.

Todos os órgãos do Poder Judiciário julgam processos. Baldan (2011) conceitua processo como um meio de garantir, por meio da constituição, o direito de uma ação, com o escopo de pacificação social, podendo ser físico ou eletrônico. Ação, para o autor, é um direito subjetivo público de requerer uma decisão sobre um pedido ao Poder Judiciário. Abreu (2016) interpreta os elementos identificadores do processo como parte, causa de pedir e pedido. Ruschel (2012) define as três fases do processo no CPC, Código de Processo Civil, de forma macro, como: postulatória, a que trata de trâmites burocráticos e cartoriais, envolvendo a petição, autuação, anexação e validação de documentos, identificação e citação das partes; saneamento e despacho saneador, que são as providências tomadas pelo juiz a fim de eliminar os vícios, irregularidades ou nulidades processuais; instrução do processo, na qual, no final, ocorre o julgamento do mérito (direito material) com a expedição da sentença. Segundo Jr (2012a), o conhecimento é o insumo

fundamental para a decisão sentencial, sendo que as atividades desempenhadas pelo magistrado são intensivas do conhecimento. Segundo o artigo 203 do novo CPC, os pronunciamentos do juiz consistem em sentenças, decisões interlocutórias e despachos. Sentença é o pronunciamento por meio do qual o juiz põe fim à fase de conhecimento do procedimento comum, bem como extingue a execução; enquanto as decisões interlocutórias são o pronunciamento judicial de natureza decisória que não se enquadre como sentença, e despachos são os demais pronunciamentos do juiz, praticados no processo, de ofício ou a requerimento da parte. Já acórdão, segundo o artigo 204 do CPC, segundo Brasil (2015), são os julgamentos colegiados proferidos pelos tribunais. Os poderes finais, ou decisórios finais, de acordo com Ruschel (2012), são os que o juiz exerce, por meio de sentenças ou atos executórios, para solucionar a lide.

2.2 Justiça Federal

A Justiça Federal, em especial a da 4^a Região, é descrita nesta subseção. A Constituição de 1988 reestruturou a Justiça Federal: foram criados cinco Tribunais Regionais Federais com sede em Brasília, Rio de Janeiro, São Paulo, Porto Alegre e Recife. A Figura 5 mostra a divisão vigente da Justiça Federal.

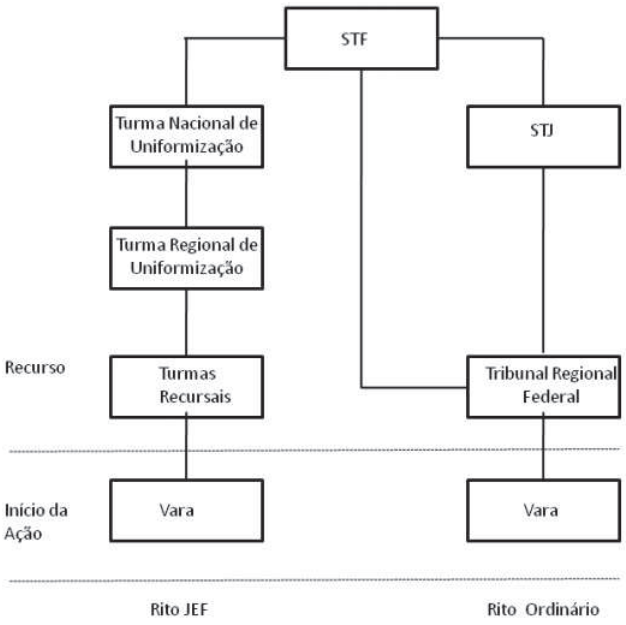
Figura 5 – Divisão geográfica da Justiça Federal Brasileira

Fonte: Conselho da Justiça Federal

Nesta tese são tratadas as matérias tributária, administrativa, previdenciária e toda e qualquer outra que não seja penal. O trâmite processual na Justiça Federal pode ser representado

pela Figura 6. O processo pode iniciar nos Juizados Especiais Federais ou, no que se denomina de 1^o Grau, dependendo do valor da causa no processo civil ou da gravidade do delito no processo criminal, e se o processo não for de competência do 1^o Grau, conforme Figura 6.

Figura 6 – Trâmite processual na Justiça Federal



Fonte: A Autora

O ingresso dos processos na Justiça Federal pode ser, segundo Esmafe-PR (2016), de forma individual ou coletiva: mesmo os conflitos de massa, ou seja, os processos que atingem um grande número de partes, como por exemplo, as ações sobre a correção monetária do FGTS, as ações previdenciárias, os processos tributários e os que tratam dos financiamentos da casa própria, podem entrar de forma individualizada.

2.2.1 Organização da Justiça Federal da 4^a Região

A Justiça Federal da 4^a Região abrange os Estados do Rio Grande do Sul, Santa Catarina e Paraná. A 2^a Instância, ou Tribunal Regional Federal da 4^a Região (TRF4), tem sede em Porto Alegre e jurisdição nos três Estados, segundo o TRF4 (2018), com as competências a seguir:

- processos em grau recursal: os recursos em causas decididas por juízes federais de primeiro grau em ações que envolvam a União Federal, autarquias e empresas públicas, recursos de decisões proferidas por juízes de direito em causas envolvendo matéria previdenciária e em execuções fiscais; e
- processos originários: os juízes federais, incluídos os da Justiça Militar e da Justiça do Trabalho, nos crimes comuns e de responsabilidade, e os membros do Ministério Público da União, ressalvada a competência da Justiça Eleitoral.

Quanto à organização, o TRF4, como os demais Tribunais, é um órgão colegiado, reunindo-se em Plenário (quando a totalidade dos Desembargadores reúne-se), Corte Especial (constituída por 15 desembargadores), Seções e Turmas. São 4 Seções integradas pelos componentes das Turmas de acordo com a área de especialização (Tributária, Administrativa, Previdenciária e Penal). As Seções Tributária, Administrativa e Penal são compostas por duas turmas, a Seção Previdenciária é composta por quatro turmas. Cada uma das turmas é composta por três Gabinetes, totalizando 30 Gabinetes, 6 de Desembargadores Federais e 6 de Juízes Federais convocados. (TRF4, 2018).

Na 4^a Região, a Justiça Federal de 1^o Grau possui 62 Subseções Judiciárias (conjunto de municípios sob a jurisdição de uma ou mais Varas Federais sediadas em um município sede). Há ainda as unidades Avançadas de Atendimento, criadas pela Lei No 10.259, de 2001 (BRASIL, 2016), que são unidades situadas em municípios que não são sede de subseção judiciária, onde são realizados os atendimentos que exijam a presença das partes processuais, como os necessários para a emissão de certidões, a realização de audiências, perícias e atermações, atendimento ao público, cadastramento de partes e advogados no processo eletrônico, qualquer ato processual que exija a atuação local de juiz ou servidor da Justiça Federal. As Varas Federais podem ter atuação exclusiva em rito ordinário, em rito JEF ou processar os dois tipos de ações, isto é, Rito JEF e Ordinário.

2.3 Principais problemas do Poder Judiciário

Retomam-se, neste capítulo, os principais problemas do judiciário brasileiro apontados no Capítulo 1, suas causas e relacionamentos, com ênfase na morosidade.

Em relação à morosidade do judiciário, um dos três principais problemas apontados pelo Ministério da Justiça, são várias as causas apontadas. Girão (2008) classifica as causas da morosidade em estruturais e relativas às ações. As causas estruturais são o elevado número de processos, o reduzido número de magistrados e servidores, a falta de gerenciamento e de investimento tecnológico, a falta de compromisso e despreparo de magistrados e serventuários. As causas relativas às ações são o formalismo processual, as mudanças de legislação e o uso indiscriminado de recursos judiciais.

Podem ser citadas outras causas da morosidade: efeitos e dificuldades que podem decorrer das etapas e garantias especificadas em lei, segundo Sadek (2004); demanda excessiva, conforme Silva (2015), sendo esta apontada por Cavalcante (2008) como a causa principal; má gestão do fluxo processual, com destaque para o desbalanceamento entre a demanda e a oferta; número reduzido de juízes, condições físicas de trabalho e número de servidores, conforme Carvalho, Gonçalves e Oliveira (2012) e Chappe (2012); e, segundo Cunha et al. (2011), a associação a uma cultura organizacional burocrática, formalista e a um modelo de gerenciamento processual ultrapassado.

A morosidade, segundo Sadek (2004), tem relação com o acesso à justiça: por um lado, há o desconhecimento de direitos de parte da população; por outro, uma justiça percebida como cara e lenta, afastando parte da população do judiciário e fazendo com que esses só procurem por justiça de forma compulsória. De acordo com Silva (2015), a morosidade desmotiva o cidadão a buscar seus direitos na justiça pela incerteza de ter seus direitos reconhecidos ou de receber uma resposta, impedindo a efetivação dos direitos fundamentais.

O segundo problema básico principal, o excesso de processos em estoque, pode ser atribuído, segundo Chappe (2012), aos recursos do tribunal, à sua eficiência e organização, e não está claro se a demanda aumenta quando a capacidade do judiciário aumenta. Então, o problema pode ser analisado tanto pela perspectiva da oferta como da demanda. O enfoque da prestação jurisdicional, segundo estudo realizado pela Pucrs (2011), diz respeito ao que se oferta como serviços para a população tais como jurisprudência, consulta processual e, principalmente, a resolução de litígios.

Do lado da demanda, de acordo com Carvalho, Gonçalves e Oliveira (2012), a procura por serviços judiciais no País aumentou nos últimos anos, resultado das taxas de industrialização e urbanização e do aumento dos conflitos sociais oriundos do crescimento das relações comerciais, conforme Sadek (2004), porém contidas pela ausência de vida democrática e pelo descrédito na justiça.

Contudo, esta demanda é bastante incerta, pois depende de fatores tais como envelhecimento da população, planos econômicos, custas processuais, entre outros. Segundo Girão (2008), os principais motivos para a alta demanda processual são o crescimento populacional, a evolução tecnológica, o exercício da cidadania, a ênfase sobre o direito das pessoas, e o êxodo rural dos anos de 1980 em função da industrialização. Feitosa (2007) cita, ainda, a inoperância dos poderes executivo e judiciário como causa do aumento da demanda. Outro fator apontado para o aumento da demanda é a garantia de direitos fundamentais por meio da promulgação da Constituição de 1988. ((BONOTTO, 2012; SILVA, 2015)).

A demanda torna-se alta porque há motivações para litigar, para recorrer e para estabelecer acordos. Há um ambiente propício à litigância à medida que os custos são baixos ou inexistentes, com baixo risco quanto ao resultado do litígio na busca por um prêmio. Da mesma forma, há motivação para o recurso à medida que a percepção da morosidade da justiça e as cumulativas possibilidades de apelo possibilitam a postergação de responsabilidades. Pucrs (2011). Este estímulo à litigância pode levar à utilização dos serviços judiciais até a exaustão. Esse cenário se assemelha ao pressuposto teórico da "tragédia dos comuns", que, segundo Hardin (1968), acontece quando cada indivíduo tem seus interesses em um mundo limitado destes recursos, levando à ruína de todos. Isto é, o acesso sem limites e a demanda sem restrições de um recurso limitado condena este mesmo recurso, esgotando-o.

A demanda elevada ainda traz problemas referentes ao desempenho: para George e Guthrie (2003) e Jonski e Mankowski (2014), a melhora no desempenho do judiciário pode

aumentar a demanda, à medida que a procura por serviços judiciais aumenta e são percebidos como céleres. Porém, segundo George e Guthrie (2003), os defensores da reforma judicial, argumentam que, com o aumento no número de juízes e tribunais, o sistema judicial será capaz de acomodar a demanda existente sem estimular nova demanda, supondo-a inelástica ou insensível a mudanças no preço implícito que os litigantes pagam ao esperar pela resolução de um caso. Para o autor, a curva do fornecimento de litigância é completamente inelástica, isto é, vertical, porque a quantidade de serviços judiciais fornecidos é constante no tempo, influenciada pela pressão de grupos de interesses e congestão nas cortes.

Além do reflexo no desempenho, de acordo com Pucrs (2011), a alta demanda implica em maior carga de trabalho, demandando alta produtividade, o que pode impactar de forma negativa na qualidade da prestação jurisdicional, refletida no índice de satisfação com os serviços prestados. Por outro lado, como conclusão do estudo realizado em tribunais da Justiça Estadual (justiça comum), os processos de trabalho não apresentaram maiores problemas de organização, e as questões mais deficientes referem-se a pessoal, à organização do trabalho, à infraestrutura e à motivação de equipes. Estes dois problemas do judiciário, a alta litigiosidade e a morosidade, de acordo com Sadek (2004), podem desgastar não só a crença da população no Poder Judiciário, mas a qualidade da democracia no País.

A respeito do acesso à justiça, de acordo com Mota (2014), o acesso está ligado à justiça social e refere-se ao fato de que todos podem postular a tutela jurisdicional preventiva ou reparatória em caso de ameaça ou defesa de direitos individuais, coletivos ou difusos. Da mesma forma, para Silva (2015), o acesso à justiça é direcionado aos mais pobres, ou seja, a interpretação da Lei deve ser direcionada a aplicá-la visando aos cidadãos pobres, garantido pelo artigo 10 da Declaração Universal dos Direitos Humanos. Desta forma, o acesso à justiça é visto como um movimento para a garantia dos direitos sociais.

Silva (2015) complementa afirmando que os com menor poder aquisitivo e menor grau de instrução são os que desconhecem seus direitos, além de os custos de litigiosidade impedirem o seu acesso à justiça; e, do lado inverso, os com maior poder aquisitivo têm dificuldades de compreensão das normas jurídicas, também dificultando seu acesso à justiça. Especificamente sobre os custos de entrada, de acordo com Buscaglia e Ulen (1997), os custos de utilizar os tribunais encorajam apenas a entrada dos casos mais complexos, aqueles que podem causar uma redução na eficiência judicial.

A partir da discussão dos problemas da morosidade e da alta quantidade de processos no país, é possível relacioná-los com o tempo de atravessamento processual. Por um lado, a elevada demanda reforça a necessidade de um processo com maior *throughput*, ou taxa de saída, que por sua vez, tem uma relação de causa e efeito com a morosidade. Isto é, a elevada distribuição processual exige uma taxa maior de saída (ou baixa de processos). Da mesma forma, tempos elevados de atravessamento levam a um menor taxa de saída, implicando maior estoque de processos e maior percepção de morosidade.

A morosidade pode ser entendida como a percepção sobre o tempo de espera pelo julgamento. Assim, dada uma certa eficiência de produção, quanto maior o número de processos em estoque, maior a morosidade; e quanto mais morosa é a justiça, maior é o estoque. Essa análise é uma ilustração da Lei de Little aplicada ao judiciário, conforme a seção 2.5.

De qualquer forma, há iniciativas para solucionar os problemas anteriormente analisados. Entre elas, podem ser citadas as metas do Conselho Nacional de Justiça, o planejamento estratégico da Justiça Federal e o novo CPC.

O Novo Código de Processo Civil traz proposta para i) diminuir a demanda, por meio da tentativa de acordo antes da abertura do processo e aplicação de multas a quem entrar com diversos recursos para adiar a decisão final; ii) reduzir a morosidade, por meio de propostas de julgamento das ações por ordem cronológica, com priorização de causas relevantes.

O planejamento estratégico da Justiça Federal tem propostas para atuar, principalmente, nos problemas de celeridade e número de processos em estoque, por meio de projetos para adoção de soluções alternativas de conflito, impulso às execuções fiscais e cíveis, gestão das demandas repetitivas e dos grandes litigantes.

As metas propostas para os Tribunais pelo CNJ pretendem atacar os três problemas: i) diminuir o número excessivo de processos, por meio das metas de julgar mais processos que os distribuídos, de aumentar os casos solucionados por conciliação, de baixar quantidade maior de processos de execução que o total de casos novos no ano corrente; ii) morosidade, por meio da meta de julgar processos mais antigos e iii) melhorar o acesso à justiça por meio da meta de priorizar o julgamento das ações coletivas.

De modo geral, todas as ações referem-se à efetividade do sistema judiciário, isto é, a sua capacidade de decidir de forma célere e justa. Neste sentido, modelar informações relacionadas ao tempo de atravessamento fornece informação sobre a morosidade, um dos três principais problemas listados.

2.4 Indicadores utilizados no Poder Judiciário

A seguir, apresentam-se dois sistemas de indicadores de referência para o sistema judiciário: um usado na Comunidade Europeia e outro, no Brasil. Antes disso, entretanto, conceitua-se o sistema *Q-Justiça*.

Segundo Serbena (2013), *Q-Justiça* é um sistema de métricas judicial (ou justiça quantitativa), complementar ao sistema E-Justiça, ou justiça eletrônica. De acordo com o autor, simultaneamente ao processo de informatização da justiça brasileira, houve a implantação de um sistema de coleta de dados e de análise estatística do Poder Judiciário e de sua performance, por meio de índices específicos.

Atualmente, a questão principal na administração judicial, segundo Hanson, Ostrom e

Kleiman (2010) é a definição e mensuração de desempenho de alta qualidade, como no setor privado. Isto é, para os autores do trabalho, as cortes devem procurar por critérios e indicadores de entrega dos serviços em padrões de produção. Por exemplo, em organizações com fins não lucrativos e em empresas públicas, de acordo com Lepore, Metallo e Agrifoglio (2012), indicadores não financeiros, tais como os presentes no BSC – *Balanced Scorecard*, podem ser usados para propor metas e verificar seu cumprimento. A seguir, são apresentados os extratos dos dois sistemas de indicadores previamente mencionados.

2.4.1 Indicadores europeus

Na justiça européia, o CPMS - *Court Performance Measure System*, definido pelo laboratório internacional para o estudo de sistemas judiciais (International Laboratory for the Study of Judicial Systems), foi baseado no BSC, e apresenta cinco dimensões, de acordo com Lepore, Metallo e Agrifoglio (2012), descritas a seguir:

- da perspectiva dos clientes: os indicadores medem a acessibilidade ao Tribunal e o tratamento aos usuários em termos de imparcialidade, equidade e respeito, mensurados mediante pesquisa de opinião;
- da perspectiva dos processos internos: são os mesmos indicadores propostos pela CEPEJ para avaliar a eficiência dos tribunais europeus: *clearance rate* (os casos resolvidos como uma porcentagem do número de casos novos), *case turnover ratio* (é a razão de casos resolvidos em relação aos casos não resolvidos), *disposition time* (365 dividido pelo indicador *case turnover ratio*); Para os autores, esses indicadores são destinados a avaliar o tempo em que um tribunal leva para processar os casos;
- da perspectiva financeira: o indicador proposto é o custo médio por processo de acordo com o tipo de processo (civil ou criminal);
- da perspectiva da inovação e aprendizagem: são incluídos os indicadores para avaliar a contribuição dos recursos humanos (número de pessoal administrativo, número de juízes, número de usuários final), informação essencial (investimento em hardware e software de tecnologia de informação e comunicação, podendo ser interpretados como uma aproximação do potencial para inovação e aprendizado), cultura do Tribunal para suportar inovação e aprendizado, mensurado por meio de questionário adaptado da Organizational Culture Assessment Instrument – OCAI com duas dimensões – solidariedade e sociabilidade. A dimensão solidariedade refere-se aos objetivos compartilhados, interesses mútuos e tarefas comuns, enquanto a sociabilidade relaciona-se ao grau em que os juízes e funcionários trabalham cooperativamente de uma forma cordial; e
- da perspectiva de sucesso dos sistemas de informação: o modelo utilizado é o de Delone e McLean, para investigar a compreensão dos sistemas de informação e seu impacto. O modelo analisa três componentes – criação, uso e consequências do uso do sistema.

De forma similar, Hanson, Ostrom e Kleiman (2010) propõem quatro perspectivas para um mapa estratégico de alta performance de tribunais: perspectiva do cliente; perspectiva das operações internas; perspectiva da inovação; e perspectiva dos valores sociais. Os autores citam em seu trabalho os indicadores do sistema CourTools:

- Acesso e justiça: tratamento dado aos usuários em termos de justiça, igualdade e respeito;
- Taxas de acesso: uma medida de acessibilidade definida como as taxas judiciais médias pagas por processo civil;
- *Clearance Rate*: os casos resolvidos como uma porcentagem do número de casos novos;
- *Disposition time*: percentual de casos resolvidos (ou não finalizados) dentro de prazos estabelecidos;
- Idade dos casos pendentes ativos: número de dias da petição até a data da medida;
- Julgamento na data agendada: a proporção de eventos que são realizados na primeira data agendada;
- Integridade e confiabilidade dos arquivos: a porcentagem de arquivos que podem ser recuperados dentro dos padrões de tempo estabelecidos, e que cumprem as normas estabelecidas para a integralidade e exatidão dos conteúdos;
- Penalidades monetárias: expressa no total de multas, taxas, restituição e custos ordenados por um tribunal;
- Efetivo uso do Júri: é a taxa na qual jurados são usados pelo menos uma vez no julgamento;
- Satisfação dos empregados: funcionários que avaliaram a qualidade do ambiente de trabalho e as relações entre o pessoal e gestão; e
- Custo médio por processo, de acordo com o tipo de processo.

Também da justiça europeia, de acordo com Kirat et al. (2010), há o Projeto de Lei do Orçamento, do programa francês *Judicial Justice*, com indicadores criados para definir *ex ante* os objetivos e avaliar os desempenhos *ex post*. Os principais indicadores são:

- Duração média dos processos de adjudicação, por nível de jurisdição;
- Porcentagem de tribunais excedendo a duração limite de processamento do caso, considerando que são definidos prazos máximos para a duração processual;
- A média do atraso em processos antigos, por tipo de jurisdição;
- Tempo médio para a entrega do julgamento com a fórmula executória;
- Taxa de anulação de sentença pelo tribunal superior em casos civis;
- Número de casos civis tratados pelo juiz;

- Número de casos tratados pelo funcionário responsável dentro dos tribunais;
- Duração média dos procedimentos criminais;
- Taxa de não-admissão do registro criminal pelo órgão *Casier Judiciarie national*;
- Taxa de anulação de sentença pelo tribunal superior em processos criminais;
- Número de casos criminais tratados pelo juiz; e
- Quantidade de crimes possíveis de serem processados por oficiais do Ministério Público.

De acordo com Kirat et al. (2010), os indicadores franceses estão no final da cadeia para avaliar o nível de desempenho na realização dos objetivos da ação do Estado. Há indicadores em dois níveis: no primeiro nível, na missão da justiça, os indicadores estão integrados no controle do Estado e o Conselho; no segundo nível, há cinco programas: justiça judicial, administração penitenciária, proteção judicial da juventude, o acesso ao direito e à justiça, administração da justiça e agências relacionadas, compreendendo 60 indicadores. Segundo Hall e Keilitz (2012), as medidas globais de performance de Tribunais (Global Measures of Court Performance), descrevem onze medidas alinhadas com os valores e áreas de excelência de cortes, respondendo às questões "O que medir" e "Como medir". Estas medidas estão alinhadas com os valores judiciais aceitos universalmente e identificados pelo IFCE - International Framework for Court Excellence. As onze medidas são:

- Satisfação dos Usuários das Cortes (acessibilidade, justiça, acurácia, temporalidade, qualidade da informação e o serviço prestado);
- Taxas de acesso (medida de acessibilidade definida como as taxas médias pagas por processo civil);
- *Case clearance rate*: taxa de casos finalizados em relação aos novos;
- Processamento *on time*: percentual de casos resolvidos (ou não finalizados) dentro de prazos estabelecidos;
- O tempo médio de réus criminais presos à espera de julgamento;
- Integridade dos registros: percentual de casos e registros que atendem aos padrões de exatidão, integridade;
- *Case Backlog*: percentagem de casos com tempo maior do que prazos estabelecidos;
- Julgamento na data agendada: a proporção de eventos que são realizados na primeira data agendada;
- Compromisso dos empregados: uma variável *proxy* para o sucesso da corte, definida como o percentual de empregados que estão produtivamente comprometidos com a missão.
- Cumprimento de ordens judiciais. Recuperação das custas judiciais criminais e civis como uma proporção das impostas (uma medida do cumprimento da legislação e da eficiência); e

- Custo por Processo: custo líquido por caso finalizado.

2.4.2 Indicadores Brasileiros

No Brasil, os indicadores do Sistema de Estatísticas do Poder Judiciário (SIESPJ), embora diferentes para os diversos segmentos de Justiça, abrangem os seguintes módulos para a Justiça Comum (Justiça Federal e Estadual), de acordo com a Resolução 76/2009 (CNJ, 2009):

- Insumos, Dotações e Graus de Utilização: com os indicadores financeiros, de recursos humanos e físicos;
- Litigiosidade, incorporando o 2^o Grau, 1^o Grau, Turma Recursal e Juizado Especial;
- Acesso à justiça;
- Tempo do processo judicial; e
- Casos novos por classe e assunto (perfil das demandas).

Esses indicadores são divulgados por Tribunal (2^o Grau). Os de litigiosidade são classificados apenas em criminais e não criminais. De forma ampla, o CNJ publica um relatório com os indicadores, entre os quais, destacam-se: IPC-JUS, taxa de congestionamento (compara o que não foi baixado com o que tramitou durante o ano); índice de produtividade por Magistrado (processos baixados por Magistrado); índice de produtividade por servidor da área judiciária; índice de atendimento à demanda (baixados em relação aos casos novos); e os tempos médios e medianos entre autuação e baixa dos processos. Em relação aos tempos, de acordo com (CNJ, 2009), o CNJ publica os seguintes indicadores, apresentando as medidas média, mediana e desvio padrão:

- Tempo de tramitação dos processos baixados;
- Tempo de tramitação dos processos pendentes;
- Tempo da decisão (terminativa, sentença);
- Tempo da suspensão (total, em função de repercussão geral, de recurso repetitivo);

A periodicidade da coleta de dados é anual, à exceção do módulo de Litigiosidade, que é semestral. Os tribunais enviam os dados ao Departamento de Pesquisas Judiciárias do CNJ, que os consolida, analisa e divulga.

O IPC-Jus, Índice de Produtividade da Justiça, conforme (CNJ, 2009), é calculado por meio de Análise Envoltória de Dados para a Justiça comum (Federal e Estadual). O índice considera como insumos, a despesa total, a quantidade de Magistrados e Servidores da área judiciária, os casos pendentes no início do ano (estoque) e casos novos (demanda). A variável considerada como produto é a quantidade de processos baixados durante o ano.

A taxa de congestionamento, calculada com base na quantidade de processos baixados e casos pendentes no final do ano, procura medir o que permanece represado na justiça anualmente, conforme (CNJ, 2009). O indicador varia de 0 a 100%; quanto maior o valor do indicador, mais congestionado está o Tribunal.

A carga de trabalho por magistrado, conforme (CNJ, 2009), é a quantidade de processos a serem trabalhados, anualmente, pelos Magistrados. Já os índices de produtividade são a quantidade de processos baixados por magistrado ou por servidor da área judiciária. O índice de atendimento à demanda, por sua vez, é a quantidade de processos baixados em relação aos casos novos. (CNJ, 2009). Um índice superior a 100% indica que a Justiça está cumprindo não apenas a demanda, mas reduzindo o estoque de processos. Por outro lado, um índice inferior a 100% indica que a Justiça não está atendendo à demanda.

Adicionalmente, Pinheiro (2008) cita os seguintes exemplos como indicadores de legitimidade, com o intuito de fortalecer o Poder Judiciário: a confiança da população no Judiciário para aferir a aproximação da sociedade ao judiciário; inserção social; e satisfação com os serviços prestados. Estes indicadores não estão contemplados por aqueles propostos pelo CNJ.

Serbena (2013) conclui que o sistema estatístico do poder judicial brasileiro está em seus passos iniciais, porém, é necessário ainda que os bancos de dados sejam publicados, em formato padronizado, com o objetivo de permitir pesquisa a diversas áreas de atuação, o que já está garantido pela Lei 12.527, a Lei de Acesso à Informação (BRASIL, 2016c).

Na Europa, o CEPEJ divulgou o relatório com estudo sobre o tempo de duração dos processos recursais dos tribunais e dos tribunais supremos dos Estados membros do Conselho da Europa (VELICOGNA, 2015). A proposta do relatório foi examinar detalhadamente a duração dos processos e o tempo necessário para processar casos pendentes com base nas informações coletadas ao longo do tempo. Com os resultados das medidas descritivas (mediana, média, mínimo e máximo) foi possível estabelecer parâmetros de tempo razoável de duração processual.

O relatório apresenta o tempo de duração dos indicadores *clearance rate* e *disposition time*. São mostradas as medidas média aritmética simples, mediana, máximo e mínimo, além da variação média anual do tempo entre dois períodos. Porém, o tempo observado não é apresentado.

Tanto na Europa quanto no Brasil, os indicadores formais de tempos processuais são coletados e divulgados de forma retrospectiva. Desta forma, acredita-se que o melhor indicador de referência para o tempo de atravessamento, constante no sistema do CNJ, seja o tempo de tramitação dos processos baixados ou duração processual.

2.4.3 Comparação entre os indicadores brasileiros e europeus

O Quadro 1 compara os indicadores utilizados no Brasil com os propostos pela *Courttool*, IFC. O relatório publicado pelo CNJ não aborda o relacionamento com os clientes ou com os funcionários, embora o CNJ faça pesquisas de satisfação com os usuários e de clima organizacional

com os funcionários e magistrados. CNJ (2011). Da mesma forma, nos indicadores brasileiros não estão presentes os indicadores de conformidade com padrões estabelecidos, sendo esta a principal falta observada em comparação aos indicadores internacionais. Os relatórios brasileiros ainda possuem indicadores de qualidade de decisão, que são os de recorribilidade e a reforma de decisão.

Quadro 1 – Comparação de indicadores usados na Europa e no Brasil

Grupo de indicador e indicador	Comunidade Europeia	Brasil
Efetividade (grupo)		
Clerance rate (taxa de casos finalizados em relação aos novos)	X	X
Disposition time (percentual de casos resolvidos ou não finalizados dentro de prazos estabelecidos)	X	X
A case *turnover ratio* (é a razão de casos resolvidos em relação aos casos não resolvidos)	X	
Duração processual		
Duração média dos processos de adjudicação, por nível de jurisdição	X	
Idade dos casos pendentes ativos	X	X
Média do atraso em processos antigos, por tipo de jurisdição	X	
Tempo médio para a entrega do julgamento com a fórmula executória	X	
Duração média dos procedimentos criminais	X	X
Conformidade		
Julgamento na data agendada	X	
Processamento *on time*	X	
Integridade e confiabilidade dos casos	X	
Integridade dos registros: percentual de casos e registros que atendem aos padrões de exatidão, integridade	X	
Porcentagem de casos com tempo maior do que prazos estabelecidos (*Case backlog*)	X	
Porcentagem de tribunais excedendo a duração limite de processamento do caso, considerando que são estipulados prazos máximos para a duração processual	X	
O tempo médio de réus criminais estão presos à espera de julgamento	X	

Continua...

Quadro 1 - Comparação de indicadores usados na Europa e no Brasil ... continuação

Grupo de indicador e indicador	Comunidade Europeia	Brasil
Penalidades monetárias	X	
Efetivo uso do Júri	X	
Satisfação dos empregados	X	
Compromisso dos empregados: uma variável Proxy para o sucesso da Corte, definida como o percentual de empregados que estão produtivamente comprometidos com a missão	X	
Taxa de anulação de sentença pelo tribunal superior em casos civis	X	
Quantidade de crimes possíveis de serem processados por oficiais do Ministério Público	X	
Número de casos criminais tratados pelo juiz	X	X
Qualidade das decisões		
Número de casos civis pelo juiz	X	X
Número de casos tratados pelo funcionário responsável dentro dos tribunais	X	X
Número de pessoal administrativo	X	X
Número de juízes	X	X
Número de usuários	X	X
Relacionamento com os usuários	X	
Imparcialidade (tratamento aos usuários)	X	
Equidade (tratamento aos usuários)	X	
Respeito (tratamento aos usuários)	X	
Taxas de acesso	X	
Satisfação dos Usuários das Cortes	X	
Financeiros		
Recuperação das custas judiciais criminais e civis como uma proporção das impostas (uma medida do cumprimento da legislação e da eficiência)	X	
Custo médio por processo	X	
Investimento em hardware e software de comunicação e tecnologia de informação	X	X

Fonte: A Autora, baseado em Kirat et al. (2010), Hanson, Ostrom e Kleiman (2010), Lepore, Metallo e Agrifoglio (2012), Hall e Keilitz (2012) e CNJ (2009)

.

Ainda, observa-se que os indicadores brasileiros não medem o custo por processo,

embora haja estudos oficiais calculando estes custos como o de Cunha et al. (2011), abordando os custos processuais em execuções fiscais. Há um esforço, tanto no Brasil como fora, em medir a eficiência da justiça: no caso brasileiro, é medido pelo IPCJUS; na Europa, embora os sistemas de indicadores sejam mais alinhados ao BSC, existem trabalhos como o de Falavigna et al. (2015) abordando a questão.

2.5 Tempo de atravessamento processual

De acordo com Tubino (2004), tempo de atravessamento é o tempo gasto para transformar matérias primas em produtos ou serviços, podendo ser considerado de forma ampla (tempo de atravessamento do cliente), quando o propósito é mensurar o tempo da solicitação até a efetiva entrega, ou de forma restrita (tempo de atravessamento de produção) com o objetivo de medir as atividades internas do sistema de produção.

Porém, segundo Ericksen, Stoflet e Suri (2007), esta definição não ajuda a compreender e eliminar os desperdícios de todo o sistema nem indica como o cumprimento da ordem de serviço é alcançada, sendo necessário um indicador com foco tanto no resultado como na forma que o resultado é alcançado: o tempo do caminho crítico da manufatura, ou seja, o tempo desde a solicitação da ordem, através do fluxo do caminho crítico, até que pelo menos um item do pedido seja entregue ao cliente. Entretanto, considerando-se as múltiplas possibilidades de recurso em um processo judicial, a primeira sentença ou decisão com o objetivo de terminar o processo em uma instância, pode ser uma peça entregue ao cliente, mas sem resolver o seu problema.

Nesta tese, será adotado o conceito de tempo de atravessamento amplo, ou seja, o tempo desde a autuação da ação judicial até o seu trânsito em julgado no 2^o grau, quando não há mais recurso possível. De acordo com Hopp e Spearman (2013), o tempo de atravessamento de uma linha de produção é uma constante determinada pela gerência, definido como o tempo disponibilizado para a produção de um item naquele roteiro ou linha.

Por outro lado, o *cycle time* de um produto é aleatório, sendo que, em um ambiente de trabalho sob encomenda, como é o caso de processos judiciais, um indicador para o desempenho é o nível de atendimento definido como

$$\text{Nível de atendimento} = \text{P}(\textit{cycle time} \leq \textit{lead time}) \tag{1}$$

Segundo Tubino (2004), é possível identificar quatro grupos de tempos que compõem o tempo de atravessamento, sendo que a soma destes definem o tempo de atravessamento produtivo:

- tempo de espera: é a soma dos tempos consumidos com a programação da produção, a espera na fila do recurso e com a espera para completar o lote, sendo proporcional ao número de etapas pelas quais o item passa, pois para cada uma delas ele sofrerá essa espera;

- tempo de processamento: tempo com a transformação da matéria-prima em produto, sendo o único que agrega valor;
- tempo de inspeção: é o tempo para a verificação do item de acordo com as especificações;
- tempo de transporte: tempo para movimentar o item entre os recursos produtivos.

Especificamente sobre o tempo de espera, de acordo com Tubino (2004), o componente de maior peso é o tempo de um item na fila de um recurso para ser processado, resultado de: desbalanceamento entre carga de trabalho e capacidade produtiva; esperas para setup e processamento dos lotes com prioridade no recurso; e problemas de qualidade no sistema produtivo.

Da mesma forma, para a Gestão Judiciária, o conhecimento do tempo de atravessamento pode ter aplicação na Lei de Little. Little e Graves (2008), citam a teoria das filas para explicar a lei, a partir da taxa de chegada dos itens ao sistema e do tempo médio de permanência no sistema para explicar o número médio de itens no sistema de filas, em um sistema estável. Entretanto, a lei é geralmente expressa utilizando a taxa de saída do sistema, para enfatizar sua aplicabilidade às operações, uma vez que o output é o atributo primário ou sua razão de ser. Esta Lei, segundo Hopp e Spearman (2013), relaciona o trabalho em processo (WIP), o *cycle time* (CT) e a taxa de produção (*throughput*, TH), na forma:

$$CT = \frac{WIP}{TH} \tag{2}$$

ou

$$TH = \frac{WIP}{CT} \tag{3}$$

Segundo Little e Graves (2008). As premissas, quando se usa a taxa de saída em vez da taxa de entrada, são:

- a conservação do fluxo, isto é, a taxa média de saída igual ou aproximadamente igual à taxa média de entrada;
- todas os trabalhos que entram no sistema, saem, isto é, os itens não se perdem no sistema;
- que o WIP tenha aproximadamente o mesmo tamanho no início e no final do intervalo de tempo;
- a idade média ou latência do WIP não é nem crescente nem declinante.

Uma propriedade relevante dessa expressão é a seguinte Hopp e Spearman (2013): o CT mantém-se constante enquanto o WIP for inferior à carga mínima que ocupa o sistema de produção. A partir desse ponto crítico, a elevação do WIP implica elevação do CT. Da mesma forma, o TH cresce enquanto este ponto crítico não é atingido, tornando-se constante a partir

deste ponto. Isso significa que CT (e seu efeito sobre o nível de serviço, por consequência), é um efeito do TH, e do WIP. Conhecendo-se o CT e o WIP pode-se estimar o TH necessário e, assim, pensar em estratégias, levando em conta os recursos, com o objetivo de diminuir a quantidade de processos em estoque ou o tempo de atravessamento, o que leva a um dos aspectos da presente tese.

2.6 Trabalhos publicados sobre o judiciário brasileiro

O objetivo desta seção é apresentar os trabalhos que utilizaram modelos quantitativos para estudo de tempo de atravessamento processual ou aplicaram mineração de dados no judiciário. Embora os trabalhos referentes a tempo de atravessamento tenham sido mencionados no Capítulo 1, tanto no Brasil como no exterior, neste momento, o foco está nas covariáveis usadas, na variável resposta, e nas técnicas utilizadas.

Silwattananusarn e Tuamsuk (2012) realizaram uma revisão da literatura sobre trabalhos empregando técnicas de mineração de dados, de 2007 a 2012, e identificaram o objetivo do uso de mineração de dados em diversas área de atuação:

- Saúde: agrupamento de categoria e atributos usados na análise de similaridades;
- Varejo: agrupar os segmentos de possíveis linhas de produtos e marcas para identificar mercado para clientes;
- Financeiro: identificar grupos de corporação de acordo com a indústria ou um segmento interno da indústria e, para cada um destes *clusters* ou agrupamentos, criar um modelo para predizer as taxas de mudança;
- Construção: agrupamento textual para descobrir grupos de acesso similar; e
- Colaboração e trabalho em equipe: na identificação de grupos de trabalhadores com tarefas.

Observa-se que os autores não identificam, nesse período e com os critérios por eles utilizados, trabalhos no contexto jurídico. Foram identificados os seguintes trabalhos utilizando mineração de dados no contexto judiciário brasileiro: Schneider (2003), Gribel (2014), Pavanelli (2008), Pavanelli (2007), Jr (2012a), Pavanelli et al. (2011), Pavanelli, Pavanelli e Costa (2013), Silva, Cunha e Talon (2013), Alcantara et al. (2014), Molinari e Tacla (2010), Chaphalkar, Iyer e Patil (2015), Ruschel (2012) e Gruginskie e Vaccaro (2018). A seguir, estes trabalhos são brevemente descritos e analisados.

O objetivo geral do trabalho de Schneider (2003) foi demonstrar a aplicação do processo de Descoberta de Conhecimento em Base de Dados, no Tribunal de Justiça do Rio Grande do Sul.

O trabalho relacionou dados de classificação processual com o tempo de tramitação (tempo de atravessamento) e com o perfil de réus em processos criminais. As seguintes relações foram analisadas:

- Seção × classe processual × especialização;
- Comarca × Seção × classe processual × especialização;
- Classificação × intervalo de tempo;
- Comarca × classificação × intervalo de tempo;
- Classificação processual × tipo de sentença;
- Classificação processual × audiência;
- Perfil do réu em processo criminal (profissão, estado civil, grau de instrução, raça e idade);
- Perfil do réu × natureza criminal (furto, estelionato e fraude, roubo e extorsão);
- Natureza × estado civil × grau de instrução × profissão × cor × idade;
- Comarca × natureza × profissão × cor × idade; e
- Comarca × natureza × profissão × cor × idade (sexo feminino).

O autor usou clustering e regras de associação para alcançar os objetivos, usando a ferramenta Weka (Waikato environment for Knowledge Analysis). Como recomendação futura, o autor citou a inclusão de dados municipais, do IBGE (Instituto Brasileiro e Geografia e Estatística) e de segurança pública.

Jr (2012a) envolveu Gestão do Conhecimento no meio jurídico, através da modelagem do conhecimento explícito (informações sobre recursos dos tribunais) e do conhecimento de especialistas no processo de decisão e sentenciamento. O autor aplicou técnicas de modelagem do conhecimento e levantou a hipótese de aplicação de técnicas de recuperação de informações tais como ontologias, técnicas de recuperação de informação e conhecimento e técnicas de extração de conhecimento; com o propósito de apoiar a decisão do magistrado. O objetivo foi de identificar, mediante entrevistas, quais informações e conhecimentos são utilizados para a tomada de decisão, relacionadas ao processo de sentença: como o magistrado busca informações para fundamentar sua decisão; como ele constroi o raciocínio para decisão; e o processo de justificação da sentença. A sugestão final do trabalho foi de que estes conhecimentos fossem modelados por ontologias e populado por extração de informações e conhecimento por meio de mineração de textos. Ou seja, uma ferramenta que apoie a busca a documentos, aumentando a assertividade do atendimento à demanda de informações solicitadas pelo magistrado, para dar apoio à atividade de fundamentação sentencial e, mais especificamente, à tarefa de justificação

da sentença, considerando o número de documentos a serem recuperados e a dificuldade na busca. O autor definiu um Sistema de Conhecimento baseado no formalismo das ontologias, por meio da recuperação de informação relevante, baseado em aplicação de metodologia para identificação de contexto e demandas / tarefas intensivas de conhecimento.

Em outro trabalho focado em rever tanto o tempo de duração das audiências como do trâmite processual, Pavanelli et al. (2011) identificaram 12 atributos específicos da justiça trabalhista, mediante entrevistas com juízes titulares:

- Rito (trabalhista ou procedimento sumaríssimo);
- Tempo de Serviço do Reclamante: (tempo em meses entre a data de dispensa e data de admissão, em meses);
- Último Salário do Reclamante;
- Profissão do reclamante (dividido em setor comércio, indústria e serviço; e cargo - direção e execução);
- Objeto do Processo (solicitações feitas pelo reclamante: falta de registro em carteira profissional, diferenças salariais, verbas rescisórias, multa do Art. 477, multa do Art. 467, horas extras e reflexos, fundo de garantia por tempo de serviço, indenização por danos morais, seguro desemprego, vale transporte, adicional de insalubridade, adicional noturno e plano de saúde);
- Juiz do órgão pesquisado (Vara Trabalhista);
- Quantidade de depoimentos em cada audiência;
- Acordo (presença ou ausência);
- Necessidade de perícia;
- Solicitação de Recurso Ordinário ao Tribunal Regional do Trabalho (TRT) contra sentença emitida pelo Juiz de 1^o Grau;
- Solicitação de Recurso de Revista contra acórdão emitido pelo TRT ao Tribunal Superior do Trabalho (TST); e
- Número de audiências até a emissão da sentença.

Como conclusão, para criar uma agenda inteligente de audiências, em vez de atribuir um tempo padrão a todas elas, os autores sugeriram o uso de técnicas de classificação e redes neurais, buscando obter o tempo necessário para a audiência. O mesmo foi sugerido para obter uma estimativa, em meses, para a duração do processo.

Gribel (2014) propôs uma abordagem para identificar possíveis resultados de julgamento, considerando a similaridade com processos já julgados, em cinco passos:

1. cálculo das medidas de similaridade a partir das variáveis relator, réu, autor, tipo, assunto, resumo e reivindicação de processos julgados pelo Supremo Tribunal Federal;
2. obtenção dos clusters a partir das medidas de similaridade;
3. a partir dos agrupamentos detectados, calcular as semelhanças entre o novo processo e os outros já classificados. Neste passo, o novo processo é atribuído ao cluster ao qual mais se assemelha;
4. compilação das decisões. Considerando, como entrada no sistema, uma lista de magistrados que vão decidir o processo e seus últimos votos; e
5. previsão da decisão: os votos anteriores são computados e o grau de concordância entre os Magistrados é calculado. O resultado provável é um número entre 0 e 1, indicando a decisão provável para o processo a ser testado.

Pavanelli, Pavanelli e Costa (2013) compararam o uso de redes neurais, regressão logística e árvores de decisão, com o objetivo de estimar a duração de um processo trabalhista. Ou seja, um trabalho semelhante ao realizado por Pavanelli et al. (2011), diferente nas técnicas utilizadas.

Já Silva, Cunha e Talon (2013) utilizaram mineração de dados para estimar o tempo de tramitação processual no Tribunal de Justiça de São Paulo. Os autores analisaram as variáveis Comarca, área do processo (trabalhista, tributária ou cível), tipo de ação (28 tipos diferentes), foro, decisão (procedência da ação, diligência cumprida, extinto sem julgamento do mérito, acordo, encerramento pelo Procon, parcial procedência da ação). Os autores pesquisaram dados no site do Tribunal de Justiça de São Paulo e utilizaram redes neurais artificiais para estimar a probabilidade das comarcas terem seus processos por faixa de tempo; por faixa de tempo de acordo com o tipo de decisão, por faixa de tempo de acordo com área processual e ainda por faixa de tempo de acordo com o tipo de decisão na comarca mais morosa. Desta forma, os autores identificaram as áreas e as comarcas que, possivelmente, poderiam alocar mais juízes.

Alcantara et al. (2014) analisaram a Constituição Federal sob a ótica da infometria, utilizando, como ferramenta de análise dos termos usados no texto da Constituição Federal, algoritmos desenvolvidos na linguagem R. A partir da análise da distribuição estatística das palavras na Constituição Federal, foi criada uma estrutura necessária ao suporte para uma análise semântica latente, de modo a facilitar o entendimento do leitor e fundamentar futuras análises semânticas. Como conclusão, os resultados apontaram para a necessidade de uma pesquisa mais abrangente para seleção de palavras, conforme sua frequência de repetição e o aprimoramento dos algoritmos de criação de listas de palavras comuns.

Molinari e Tacla (2010) criaram uma ontologia jurisprudencial e um aplicativo para a sua gestão, com a capacidade de extrair conceitos e propriedades a partir de documentos de acórdãos, populando-os na ontologia, através da criação de indivíduos. Os autores sugerem, para trabalhos futuros, partindo do aplicativo proposto e da ontologia por eles populada, uma

busca semântica jurisprudencial com o objetivo de obter resultados mais precisos do que aqueles obtidos atualmente nas pesquisas disponibilizadas pelos tribunais.

Ruschel (2012) estudou o desenvolvimento de um modelo para possibilitar a representação da organização do conhecimento constante na análise do direito processual na audiência de instrução e julgamento da Justiça Trabalhista. Para tanto, o autor utilizou Engenharia de Conhecimento. O modelo de conhecimento desenvolvido tem como objetivo ajudar a processar e julgar mais processos, dando qualidade às análises e decisões do juiz, porém, a efetividade do modelo de conhecimento, caso ele for implementado em um programa de computador e utilizado por uma vara trabalhista, só poderá ser avaliada a médio e longo prazo.

Gruginskie e Vaccaro (2018) ajustaram um modelo ao tempo de atravessamento processual da Justiça Federal utilizando a variável resposta agrupada em faixas de tempo. Os autores compararam o desempenho de quatro modelos quanto à acurácia, à precisão, à sensibilidade, à especificidade e F1. Os modelos foram ajustados usando máquina de vetor suporte, florestas aleatórias, redes neurais e Naive Bayes. Como conclusão, o modelo ajustado por máquina de vetor suporte mostrou os melhores resultados.

O objetivo de Schneider (2003) foi aplicar KDD, descoberta de conhecimento em banco de dados, ao judiciário, especificamente, à Justiça Estadual de 1^a instância, para encontrar padrões, relacionando dados de classificação processual (classe, seção e especialização) com o padrão de sentença, com o tempo de tramitação, com a comarca e com a ocorrência de audiência. O trabalho foi pioneiro, com resultados interessantes para a época, fornecendo informações para a Gestão Judiciária, tais como onde os processos costumam ser mais frequentes e onde são mais morosos.

Os trabalhos de Pavanelli et al. (2011) e Pavanelli, Pavanelli e Costa (2013) são continuações aos trabalhos de Pavanelli (2007) e Pavanelli (2008), implementando as sugestões a trabalhos futuros dadas pelos autores. A sugestão que ainda não foi implementada é a de aplicar Rede de Bayes para estimar o tempo de duração de audiências. O artigo de Pavanelli et al. (2011), embora não tenha o objetivo de prever o tempo de audiência, inclui a árvore de decisão nas técnicas de comparação utilizadas para estimar o tempo de atravessamento.

O estudo de Rosa (2017) teve como objetivo construir um modelo para o tempo de tramitação dos processos dos Juizados Especiais Cíveis da Justiça Estadual de Santa Catarina, identificando os principais fatores de causa e efeito que afetam o tempo de tramitação. Ao final, o autor propôs um modelo estatístico como ferramenta auxiliar de análise na busca da redução do tempo de tramitação dos processos nos Juizados Especiais Cíveis do Estado.

Comparativamente aos trabalhos anteriormente mencionados, e como apresentado no Capítulo 1, a contribuição desta tese se dá à medida em que apresenta um modelo conceitual para a geração de informações sobre o tempo de tramitação. Há, portanto, similaridades em propósito com alguns dos trabalhos, mas diferenças em estrutura, a exceção do artigo de Gruginskie e

Vaccaro (2018). Em relação às variáveis utilizadas, as usadas nesta tese são diferentes, pelo menos parcialmente, porque se trata de um tipo de justiça distinto. Da mesma forma, os bancos de dados são diferentes no sentido de que usam sistemas diferentes, codificações diferentes para assuntos, classes, e consequentemente, variáveis distintas.

Então, diferente dos trabalhos anteriores, esta tese aborda: i) os modelos de previsão relativos à Justiça Federal, ii) a comparação dos resultados dos modelos por meio de análise de sobrevivência, iii) modelos utilizando o tempo de atravessamento em faixas de tempo e iv) apresenta uma revisão de trabalhos sobre previsão de tempo de atravessamento no judiciário brasileiro.

Observa-se, nos trabalhos anteriores, uma diversidade de tamanho dos bancos de dados usados. Schneider (2003) trabalhou com uma quantidade grande de dados (77.741 registros) da justiça, enfatizada naquele trabalho. Já Silva, Cunha e Talon (2013) utilizou uma base de 5.740 processos consultados, via Internet, nos sites dos tribunais. Pavanelli (2007), Pavanelli et al. (2011) e Pavanelli, Pavanelli e Costa (2013) utilizaram bases de 100 processos cada um. Pavanelli (2008) utilizou uma base de 108 processos, enquanto o trabalho de Gribel (2014) utilizou 16 processos, consultados manualmente e incluindo variáveis consultadas no próprio processo. A diversidade de bancos de dados e número de casos mostra, também, que os resultados podem apresentar níveis diferentes de qualidade descritiva e preditiva (acurácia, RMSE). Tipicamente, ainda que as técnicas usadas sejam similares e oriundas da mineração de dados, os trabalhos de Schneider (2003) e Silva, Cunha e Talon (2013) aproximam-se mais do esperado neste contexto de análise e da aplicação dessa abordagem (PROVOST; FAWCETT, 2013).

Desta forma, esta tese aproxima-se aos trabalhos de Pavanelli et al. (2011), por propor um modelo que prevê o tempo de atravessamento processual e ao trabalho de Gruginskie e Vaccaro (2018), pela análise do tempo em faixas.

Quanto às técnicas de análise, tanto o trabalho de Pavanelli (2007) quanto o de Silva, Cunha e Talon (2013) utilizaram redes neurais artificiais para estimar o tempo de atravessamento e os autores concluíram que a técnica apresenta bom desempenho para tal. O trabalho de Pavanelli, Pavanelli e Costa (2013) compara as técnicas de redes neurais, análise de regressão e árvore de decisão para a estimação do tempo de atravessamento.

A partir das sugestões para trabalhos futuros, a que será seguida, em parte, é a de Silva, Cunha e Talon (2013), sugerindo uma maior quantidade de dados para análise. A sugestão de Schneider (2003), de incluir dados municipais, embora interessante, não será cumprida por que se acredita que seria pouco relevante para a Justiça Federal. Já as sugestões de Pavanelli (2007) e Pavanelli (2008) já foram implementadas pelos autores. O que poderá ser utilizado é o que foi feito por Pavanelli (2007), a aplicação de análise de componentes principais (ACP) às 32 variáveis. O autor teve como objetivo avaliar a importância relativa das variáveis que compõem a amostra de dados. Mas, a ACP também pode ser útil para a transformação de variáveis correlacionadas em componentes não correlacionados. Por este motivo, essa abordagem

também é explorada.

Ao finalizar este capítulo, o Quadro 2 apresenta uma sistematização das informações relevantes para resumir os trabalhos produzidos. O próximo capítulo continua o referencial teórico.

Quadro 2 – Trabalhos realizados sobre modelagem de tempo de atravessamento no judiciário brasileiro

Referência	Contexto de Aplicação	Objetivos	Variáveis	Técnicas	Sugestão de trabalhos futuros
Schneider (2003)	Justiça Estadual do RS	Identificar relações entre o tempo e diversas variáveis	Comarca, Seção, classe processual, especialização, classificação, intervalo de tempo, classificação processual, tipo de sentença, audiência, perfil do réu em processo criminal (sexo, profissão, estado civil, grau de instrução, raça e idade), natureza criminal (furto, estelionato e fraude, roubo e extorsão)	Cluster e regra de associação	Incluir dados municipais e de segurança pública
Pavanelli (2007)	Justiça Trabalhista do PR (Comarca de São José dos Pinhais)	Prever o tempo de duração de um processo trabalhista	Faixa salarial, faixa de tempo de serviço, profissão, número de audiências e as variáveis binárias (falta de registro em carteira, horas extras, fundo de garantia por tempo de serviço, verbas rescisórias, seguro desemprego, vale transporte, adicional de insalubridade, multa (Artigo 477), adicional noturno, diferenças salariais, multa (Artigo 467) e danos morais e tempo processual	Análise de componentes principais, redes neurais artificiais	Utilização de outra ferramenta como redes bayesianas

Continua...

Quadro 2 - Trabalhos realizados sobre modelagem de tempo de atravessamento no judiciário brasileiro ... continuação

Referência	Contexto de Aplicação	Objetivos	Variáveis	Técnicas	Sugestão de trabalhos futuros
Pavanelli (2008)	Justiça Trabalhista do PR (Comarca de São José dos Pinhais)	Estimar o tempo aproximado de cada audiência	Rito, tempo de serviço, salário do reclamante, profissão, faixa salarial, as variáveis binárias relativas ao objeto do processo (plano de saúde, adicional noturno, falta de registro em carteira, horas extras, fundo de garantia por tempo de serviço, verbas rescisórias, seguro desemprego, vale transporte, adicional de insalubridade, multa (Artigo 477), diferenças salariais, multa (Artigo 467) e danos morais), Juiz, presença de acordo, quantidade de depoimentos e tempo de audiência	Regressão linear múltipla e redes reurais	Utilização de outra ferramenta como análise de regressão

Continua...

Quadro 2 - Trabalhos realizados sobre modelagem de tempo de atravessamento no judiciário brasileiro ... continuação

Referência	Contexto de Aplicação	Objetivos	Variáveis	Técnicas	Sugestão de trabalhos futuros
Silva, Cunha e Talon (2013)	Justiça Estadual de SP	Identificar os tipos de processos com maior número e comarcas	Número processo, ação, área, comarca, tempo de tramitação, foro e decisão		Desenvolver uma ferramenta para a coleta dos dados para obter uma maior quantidade de dados (procedência da ação, diligência cumprida, extinto sem julgamento do mérito, acordo, encerramento procon, parcial procedência da ação)
Pavanelli et al. (2011)	Justiça Trabalhista do PR (Comarca de São José dos Pinhais)	Prever o tempo de duração de audiências trabalhistas e o tempo de duração do trâmite	Rito, tempo de serviço, salário do reclamante, profissão, objeto do processo, Juiz, depoimentos, acordo, perícia, recurso ordinário, recurso de revista, número de audiências	Análise de componentes principais, redes neurais artificiais e regressão linear múltipla	Inclusão de processos mais recentes na base de dados com o objetivo de obter um maior dinamismo e acurácia do sistema judiciário

Continua...

Quadro 2 - Trabalhos realizados sobre modelagem de tempo de atravessamento no judiciário brasileiro ... continuação

Referência	Contexto de Aplicação	Objetivos	Variáveis	Técnicas	Sugestão de trabalhos futuros
Pavanelli, Pavanelli e Costa (2013)	Justiça Trabalhista do PR (Comarca de São José dos Pinhais)	Comparar o desempenho das técnicas redes neurais artificiais, regressão múltipla e árvores de decisão na previsão do tempo de duração do trâmite de processos trabalhistas	Rito, tempo de serviço, salário do reclamante, profissão, objeto do processo, acordo, perícia, recurso ordinário, recurso de revista, número de audiências	Análise de componentes principais, redes neurais artificiais, regressão linear múltipla e árvore de decisão	Inclusão de processos mais recentes na base de dados com o objetivo de obter um maior dinamismo e acurácia do sistema judiciário, que é a mesma sugestão anterior
Gribel (2014)	Supremo Tribunal Federal	Propõe uma abordagem para identificar resultados de julgamentos	Relator, reivindicação, autor, réu, tipo, assunto, resumo e decisão	Análise de *Clusters*	Utilizar outras técnicas de aprendizado de máquina, outras técnicas de agrupamento diferentes da hierárquica e diferente número de cluster, simular e analisar decisões monocrática separadamente dos recursos internos

Continua...

Quadro 2 - Trabalhos realizados sobre modelagem de tempo de atravessamento no judiciário brasileiro ... continuação

Referência	Contexto de Aplicação	Objetivos	Variáveis	Técnicas	Sugestão de trabalhos futuros
Rosa et al. (2017)	Justiça Estadual de Santa Catarina	Propôs modelo para o tempo de tramitação, identificando empiricamente os principais fatores de causa e efeito que afetam o tempo de tramitação	Número do processo, foro, entrância, subseção, região, vara, data da distribuição, assunto, situação, parte ativa, parte passiva, data da sentença, tipo da sentença, data da última movimentação, valor da causa, quantidade de movimentos processuais	Modelo liner hierárquico	A inclusão de varáveis políticas, sociais e econômicas da cidade no qual está situada a Comarca. A inserção de variáveis da Comarca relacionada aos Recursos Humanos do PJSC
Gruginskie e Vaccaro (2018)	Justiça Federal da 4ª Região	Compararam quatro modelos de predição de faixa de tempo de tempo de atravessamento processual de acordo com variáveis qualitativa	Classe, assunto, entidade, gabinete, meio processual, origem do processo	Redes neurais, florestas aleatórias, naive Bayes e máquina de vetor suporte	Incluir variáveis quantitativas, analisar o tempo de atravessamento também em outras instâncias e analisar, tendo como variável resposta, uma variável quantitativa

Fonte: A Autora

2.7 Considerações sobre o capítulo

O capítulo contextualizou a situação do sistema judiciário brasileiro frente ao problema e objetivo propostos nesta tese. Compreender a organização do Poder Judiciário brasileiro permite evidenciar sua complexidade e as razões pelas quais este problema se justifica no contexto de um estudo de doutoramento. O propósito de abordar a organização da Justiça Federal, sua história e, principalmente, sua competência, possibilita compreender a delimitação dos resultados da análise futura.

Tratar os principais problemas do Poder Judiciário e, principalmente, as suas relações, além de mostrar a necessidade de gestão de produção no judiciário, tenta justificar a premissa de que produzir informação sobre o tempo de atravessamento serve de subsídio para a gestão da produção e ainda, na tomada de decisão das partes processuais, impactando a demanda.

O propósito de abordar sucintamente o tempo de atravessamento foi alinhar os conceitos e assim servir de apoio para a compreensão da tese. Quanto aos indicadores, o propósito de sua abordagem foi mostrar que, embora alinhados ao BSC e possibilitando medir a eficiência das Cortes, ainda há uma carência de indicadores prospectivos.

3 TÉCNICAS DE MINERAÇÃO DE DADOS E ANÁLISE DE SOBREVIVÊNCIA

Neste capítulo, são apresentados o conceito de ciência de dados, os principais conceitos de mineração de dados: o termo usado, a história, o modelo CRISP, as técnicas de modelagem de máquina de vetor suporte para classificação e regressão, redes neurais para regressão e análise de sobrevivência. Por último, são abordadas as regras de associação.

3.1 Ciência de dados

O termo ciência de dados surgiu da fusão de conceitos de estatística e de mineração de dados, conforme Hamilton (2013), sendo introduzido como uma disciplina independente em 2001, embora já tenha aparecido nas décadas de 1960, 1970 e 1980.

Para Hamilton (2013), ciência de dados é a arte de transformar dados em ações por meio da criação de produtos de dados, que fornecem informações sem a necessidade de análises subjacentes para os tomadores de decisão. De acordo com Zumel, Mount e Porzak (2014), a ciência de dados é um esforço colaborativo que planeja um número de papeis, ferramentas e habilidades. Segundo os mesmos autores, estes papeis são: o patrocinador do projeto, representando o interesse do negócio e considerado o papel mais importante; o cliente, representando o interesse do usuário; o cientista de dados, aquele que executa a estratégia analítica, comunicando-se com o patrocinador e o cliente; o arquiteto dos dados, aquele que gerencia os dados e o seu armazenamento; e o operacional, o que gerencia a estrutura.

A Figura 7 apresenta o ciclo de vida de um projeto de ciência de dados. A descrição de cada estágio é apresentada a seguir:

Figura 7 – Ciclo de vida de um projeto de ciência de dados

Fonte: Zumel, Mount e Porzak (2014)

- Definição da meta: refere-se à definição de uma meta mensurável, gerando critérios de parada e aceitação concretos.
- Coleta e gerenciamento dos dados: refere-se à identificação dos dados necessários, sua exploração e as condições para torná-los adequados à análise, sendo o passo que consome maior tempo. Diz respeito à disponibilidade, qualidade, suficiência;
- Modelagem: é o estágio de tentar extrair informações úteis a partir dos dados para atingir os objetivos, através de técnicas estatísticas ou de aprendizagem de máquina. As tarefas mais comuns neste estágio são: classificação, *scoring, ranking*, agrupamento, correlação e caracterização;
- Avaliação e crítica do modelo: é a avaliação do modelo quanto à acurácia, generalização, o desempenho, a comparação com o modelo corrente, e se os resultados (coeficientes, regras, agrupamentos) fazem sentido no contexto do problema;
- Apresentação e documentação: é a apresentação dos resultados para o patrocinador do projeto e outros *stakeholders*. É também a documentação do modelo para a organização e para os responsáveis pelo uso e execução do modelo. A apresentação engloba como interpretar o modelo, as saídas e outras informações pertinentes; e

- Manutenção do modelo: quando o modelo é colocado em operação, deve-se assegurar que ele vai funcionar sem problemas e se adaptará às mudanças do ambiente. Nesta fase, ainda se pode ajustar o modelo, caso os testes apontem esta necessidade.

Conforme a Figura 7, há relações de realimentação entre diversos estágios apresentados. Essas relações indicam que o processo de desenvolvimento de um projeto de ciência de dados é o produto de refinamentos sucessivos entre etapas adjacentes, até que se obtenham os resultados esperados e com a assertividade esperada.

3.2 Mineração de dados

Esta seção apresenta a definição de mineração de dados, os componentes, a abordagem CRISP-DM para a mineração de dados, as abordagens de máquina de vetor suporte para classificação e regressão e redes neurais para classificação.

3.2.1 Mineração de dados e aprendizado de máquina

O termo mineração de dados, segundo Han e Kamber (2006) e Han, Kamber e Pei (2011), é usado como sinônimo de KDD - *Knowledge Discovery from Data*, descoberta de conhecimento em banco de dados, e está atualmente equivocado, sendo o mais apropriado mineração de conhecimento a partir de dados, embora mineração de dados ainda seja considerada uma etapa no processo de descoberta de conhecimento.

Muitos outros termos apresentam significado similar ou próximo, tais como *knowledge mining from data* (extração de conhecimento a partir de dados), *knowledge extraction* (extração de conhecimento), *data/pattern analysis* (análise de dados/padrões), *data archaeology* (arqueologia de dados), e *data dredging* (dragagem de dados).

A mineração de dados pode, segundo Han e Kamber (2006), ser vista como o resultado da evolução da tecnologia da informação a partir da administração dos bancos de dados no funcionamento de várias funcionalidades críticas. Estas funcionalidades incluem a coleta de dados, criação de bancos, gerenciamento e análise avançada, podendo a mineração de dados ser considerada, segundo Silwattananusarn e Tuamsuk (2012), um subcampo na gestão do conhecimento.

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996) e Rajaraman et al. (2012), os estatísticos foram os primeiros a usar o termo mineração de dados, originariamente depreciativo, referindo-se a tentativas de extrair informação não suportada pelos dados. Hoje o termo assumiu um significado positivo, como a construção de um modelo estatístico.

O conceito para mineração de dados, dado por Provost e Fawcett (2013), é a habilidade que envolve a aplicação não só de tecnologia, mas ciência e arte, que estrutura o problema e permite razoável consistência, repetibilidade e objetividade. Já Fayyad, Piatetsky-Shapiro e

Smyth (1996) conceituam mineração de dados como a descoberta de novos padrões em bancos de dados, concentrando-se em algoritmos para extrair conhecimento útil. Hall, Witten e Frank (2011) complementam a definição citando a quantidade substancial de dados, através de um processo automático ou semi automático, enfatizando que os padrões descobertos, geralmente, levam a alguma vantagem de natureza econômica.

Rajaraman et al. (2012) e Han e Kamber (2006) definem como mineração de dados a descoberta de modelos para os dados. Para Han e Kamber (2006), a descoberta de padrões pode ser usada para a tomada de decisão, controle do processo, gestão da informação e processamento de *querys*, e é interessante se:

- é facilmente compreendida pelos humanos;
- é válida em uma nova base de dados com algum grau de certeza;
- é potencialmente útil;
- valida uma hipótese, isto é, um padrão interessante que representa conhecimento.

Quanto aos algoritmos, segundo Zaki e Jr (2006), mineração de dados compreende os algoritmos básicos que permitem obter *insights* e conhecimentos fundamentais de uma grande massa de dados. Desta forma, de acordo com Rajaraman et al. (2012), cientistas da computação têm olhado para mineração de dados como um problema de algoritmos, onde o modelo de dados é a resposta a uma consulta sobre o mesmo. Para Han e Kamber (2006) e Zaki e Jr (2006), mineração de dados vai além dessa visão e envolve uma integração de técnicas de diversas disciplinas como banco de dados e *datawarehouse*, estatística, aprendizado de máquina, computação de alta performance, reconhecimento de padrões, redes neurais, visualização de dados, recuperação de informação, imagem e processamento de sinais, análise de dados temporal e espacial. A Figura 8 apresenta esta visão.

Figura 8 – Mineração de dados como uma junção de múltiplas disciplinas

Fonte: Han e Kamber (2006)

Entre as disciplinas que compõem a mineração de dados, aprendizado de máquina deve ser destacada. Segundo Provost e Fawcett (2013), a mineração de dados começou como um ramo de aprendizado de máquina e os termos ainda estão ligados: ambos referem-se à análise de dados para encontrar padrões úteis e informativos, compartihando técnicas e algoritmos, com as comunidades científicas também ligadas.

Porém, mineração de dados e aprendizado de máquina apresentam diferenças: para Rajaraman et al. (2012), a mineração de dados é considerada como sinônimo de aprendizado de máquina porque usa algoritmos com um conjunto de treinamento tais como rede de Bayes, máquinas de vetor suporte e árvores de decisão.

De acordo com Provost e Fawcett (2013), aprendizado de máquina refere-se a vários tipos de melhora de desempenho, incluindo outros subcampos como robótica, que não fazem parte de KDD e mineração de dados, relacionando-se, por vezes, a assuntos como agência e cognição, que também não fazem parte do escopo da mineração de dados.

Para Han e Kamber (2006), um sistema de análise de dados que não possui uma grande quantidade de métodos é melhor classificado como sistema de aprendizado de máquina, ou seja, um sistema de análise de dados baseado em aprendizado sobre dados e estatística.

O caso típico quando aprendizado de máquina é uma boa abordagem, de acordo com Rajaraman et al. (2012) é quando há pouca noção do que se está procurando nos dados, como por exemplo, no "desafio Netflix", quando os algoritmos prevêm as avaliações de filmes por usuários baseados em uma amostra de suas respostas. Porém, de acordo com os autores, quando é possível descrever os objetivos da mineração mais diretamente, o aprendizado de máquina não se mostrou bem-sucedido.

Um outro novo conceito é *deep learning*. Segundo Lewis (2016), *deep learning* é uma área de aprendizado de máquina emergente da interseção de redes neurais, inteligência artificial, modelagem gráfica, otimização, reconhecimento de padrões e processamento de sinais. Abrange tanto o aprendizado supervisionado como o não supervisionado, utilizando modelos de aprendizado de máquina de múltiplas camadas.

De acordo com Lewis (2016), o processo básico de aprendizagem em *deep learning* pode ser dividido em quatro componentes:

- Armazenagem de dados;
- Abstração ou a tradução dos dados armazenados em representação e conceitos;
- Generalização, que usa dados abstratos para criar conhecimento e inferências em direção a novos contextos;
- Avaliação, a qual fornece o feedback para medir a utilidade do conhecimento aprendido e informar possiveis melhorias.

Na prática, de acordo com Lewis (2016), há cinco passos para o desenvolvimento de um algoritmo de *deep learning*:

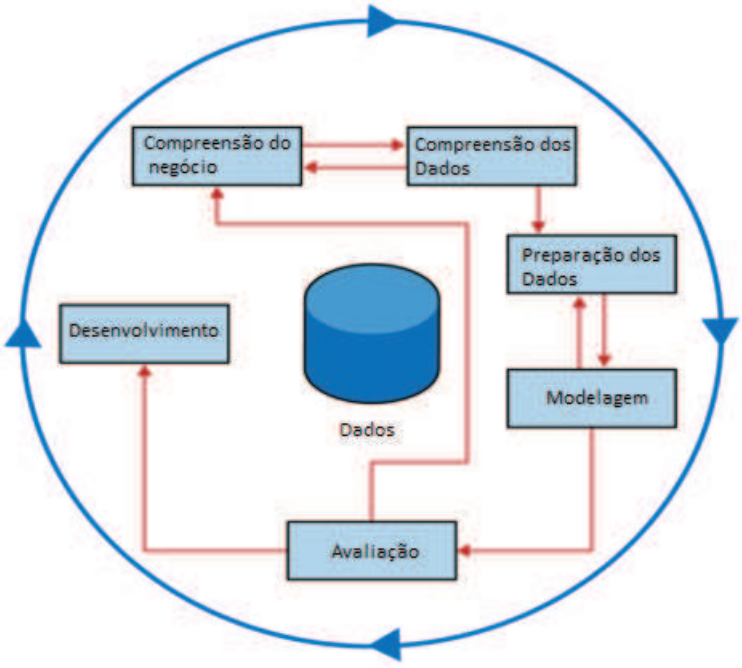
- Coleta dos dados;
- Exploração e preparação dos dados;
- Treinamento do modelo;
- Avaliação;
- Melhoria do modelo através da escolha de um diferente tipo de modelo, ou o uso de dados adicionais.

Como os conceitos apresentam passos semelhantes, será descrita a abordagem da mineração de dados. As próximas seções apresentam outros tópicos sobre mineração de dados como o modelo CRISP-DM, os componentes de mineração e análise dos dados.

3.2.2 *Cross-Industry Standard Process for Data Mining*

De acordo com North (2012), em 1999, várias empresas como a Daimler-Benz, o provedor de seguros OHRA, a fabricante de hardware e software NCR Corp. e a fabricante de software estatístico SPSS Inc. começaram a trabalhar conjuntamente para normalizar e padronizar uma abordagem para a mineração de dados. O resultado foi o *Cross-Industry Standard Process for Data Mining* (CRISP-DM), representado na Figura 9.

Figura 9 – O processo de Data Mining



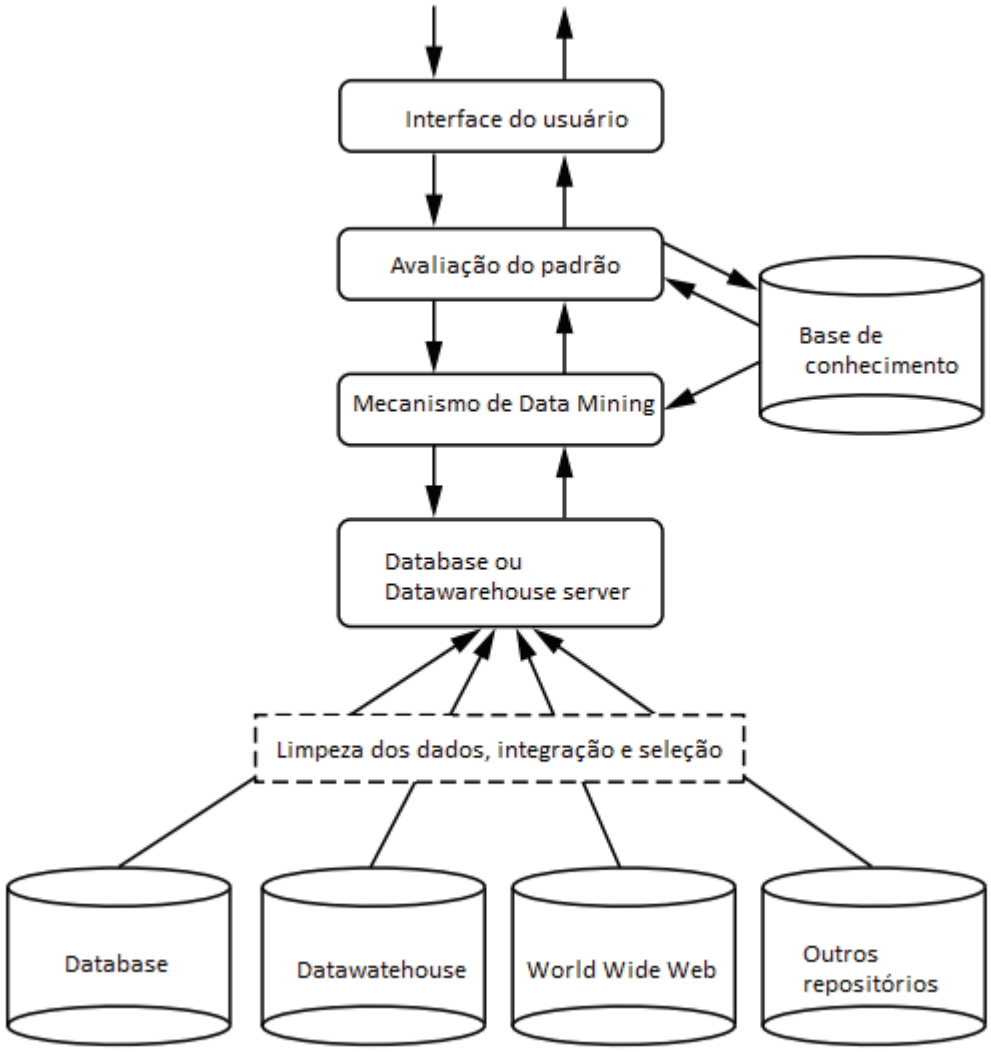
Fonte:North (2012)

Os passos do modelo conceitual são comentados no Capítulo 4.

3.2.3 Componentes de um sistema de mineração de dados

Quanto à arquitetura, segundo Han e Kamber (2006), um sistema típico de mineração de dados deve ter os seguintes componentes (Figura 10):

Figura 10 – Arquitetura de um sistema típico de mineração de dados



Fonte: Han e Kamber (2006)

- *Database* (banco de dados), *datawarehouse* (armazém de dados), *World Wide Web* ou outros repositórios, sobre os quais técnicas que visem à limpeza e à integração de dados podem ser realizadas. Ainda, da perspectiva do *datawrehouse*, a mineração de dados pode ser vista como um estágio avançado de processamento analítico *online* (OLAP) (HAN; KAMBER, 2006);
- *Database* ou *datawarehouse server*: o responsável por buscar os dados relevantes de acordo com os objetivos do usuário;

- *Knowledge base*: o conhecimento de domínio usado para orientar a pesquisa ou avaliar os padrões resultantes;
- *Mecanismo de mineração*: consiste de um conjunto de módulos para tarefas como caracterização, análise de correlação e associação; classificação, predição, análise de *cluster*, análise de *outliers* e análise de evolução;
- *Padrão de avaliação*: componente que emprega medidas de interesse (indicadores de qualidade), interagindo com o módulo de mineração de dados, com o objetivo de concentrar a pesquisa por padrões de interesse. Também pode usar limiares de interesse para filtrar os padrões descobertos e, ainda, o módulo de avaliação padrão pode ser integrado com o módulo de mineração. Segundo Han e Kamber (2006), há várias medidas de interesse baseadas na estrutura de padrões descobertos e nas estatísticas subjacentes; e
- Interface de uso: o módulo que faz a comunicação entre os usuários e o sistema de mineração de dados, permitindo a interação com o sistema e o usuário, fornecendo informações para a mineração de dados exploratória baseada na mineração intermediária.

Para Han e Kamber (2006), a mineração de dados pode ser aplicada a qualquer tipo de dado de interesse para um objetivo de aplicação, sendo as formas básicas o banco de dados, armazém de dados e dados transacionais. Pode ainda ser aplicada a outras formas como dados sequenciais, fluxo de dados, dados gráficos ou em rede, dados espaciais, texto, multimídia e dados provenientes da internet. As fontes de dados podem ser banco de dados, *data warehouses*, web, outros repositórios de informação ou dados que são transmitidos para o sistema dinamicamente. O Quadro 3 conceitua os repositórios de dados a partir dos quais a mineração pode ser realizada.

Quadro 3 – Repositórios de dados a partir dos quais a mineração pode ser realizada

Sistema de Banco de Dados	Também conhecido como *database management system*, consiste de uma coleção de dados inter-relacionados e um conjunto de programas para os gerenciar e os acessar. Estes programas permitem definir a estrutura e armazenagem, o compartilhamento, o acesso e assegurar consistência e seguridade da informação.
Data Warehouse	É um repositório de informação coletado de diversas fontes, armazenado em uma estrutura unificada, geralmente localizado em um único local. Para a construção de um *data warehouse*, são seguidos os seguintes passos: limpeza dos dados, integração, transformação, carregamento e atualização.
Banco de dados transacionais	Cada registro em um banco de dados transacional captura uma transação, que inclui um número único de identidade *t* (trans ID) e uma lista dos itens que compõem a operação. Pode ter tabelas adicionais contendo outras informações relacionadas.

Fonte: Han e Kamber (2006)

3.2.4 Análise de Dados

Os seguintes padrões de análise são identificados, de acordo com Han e Kamber (2006): descrição por conceito/classe, análise de associações e correlações, classificação e predição, análise de *cluster*, análise de *outliers*, análise de evolução. De acordo com a abordagem de mineração de dados utilizada, técnicas de outras disciplinas podem ser aplicadas como redes neurais, análise *fuzzy*, representação de conhecimento e computação de alto desempenho, análise de dados espaciais, recuperação de informação, reconhecimento de padrões, análise de imagens, processamento de sinais, computação gráfica, tecnologia da web, análises econômicas, de bioinformática ou de psicologia. Como observado, há muitas abordagens para modelar dados. De acordo com Rajaraman et al. (2012), sumarização e extração das características mais proeminentes do banco de dados integram essas opções. A seguir, apresenta-se uma breve explicação de análise exploratória de dados, mineração de padrões frequentes, agrupamento (*clustering*), classificação, regressão, redes neurais, dentre outras técnicas:

- Análise de Dados Exploratória: segundo Zaki e Jr (2006), o objetivo é explorar os atributos numéricos e categóricos de dados univariados ou multivariados, para extrair as principais características através de medidas de dispersão e tendência central. Outro objetivo da análise exploratória é reduzir a quantidade de dados a ser minerada.
- Mineração de padrões frequentes: conforme Zaki e Jr (2006), refere-se à extração de padrões úteis e informativos de uma massa de dados complexa. Padrões, neste caso, compreendem conjuntos de itens ou padrões complexos, como sequências e gráficos que consideram relações arbitrárias entre pontos, com o objetivo de descobrir tendências e comportamentos ocultos nos dados para compreender as interações entre os pontos e atributos.
- Agrupamento ou *clustering*: segundo Zaki e Jr (2006), é a tarefa de dividir os pontos em grupos naturais, os *clusters*, de acordo com a similaridade entre os pontos, considerando que estas categorias não estão classificadas previamente. Porém, de acordo com Kirk (2014), o agrupamento tem um aspecto negativo, o chamado teorema da impossibilidade, que é a impossibilidade de ter mais de dois atributos de riqueza (a existência de uma função de distância que gera todos os diferentes tipos de agrupamentos), invariância de escala (mudando a escala de medida, os agrupamentos permanecem os mesmos) e consistência (se a distância entre os pontos dentro de um aglomerado é expandida, o aglomerado deve produzir o mesmo resultado.)
- Classificação: formalmente, é uma função que prediz a classe para um novo dado com base em dados anteriores. Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), classificação é uma função que categoriza os itens dos dados em uma das classificações pré-definidas e, segundo Zaki e Jr (2006), a tarefa de classificação é predizer a classe para um ponto não classificado. Para a construção da função, há um conjuntos de dados (de treinamento) em

que há pontos classificados e após aprender o modelo, automaticamente, pode-se predizer a classe para outro novo ponto. As técnicas de classificação podem ser probabilísticas, máquinas de vetor suporte e árvores de decisão, entre outras.

- Regressão: de acordo com Kirk (2014), a ideia da regressão é ajustar uma linha de predição para uma ou mais variáveis de resposta com base em dados de uma ou mais variáveis preditoras.
- Regressão logística: de acordo com Zumel, Mount e Porzak (2014) é apropriada quando o objetivo é estimar a probabilidade de um objeto ser classificado em determinada classe, sendo uma boa escolha quando o objetivo é ter uma ideia do impacto relativo de diferentes variáveis de entrada na variável de saída.
- Classificação simples de Bayes: segundo Kirk (2014), é um método de aprendizado supervisionado e probabilístico, em que os dados de entrada são independentes um dos outros. De acordo com Zumel, Mount e Porzak (2014), é especificamente útil para problemas com muitas variáveis de entrada, ou variáveis categóricas com muitas categorias e ainda classificação de texto, sendo uma primeira tentativa de resolver o problema da categorização.
- Redes neurais artificiais: segundo Han e Kamber (2006), redes neurais artificiais têm sua origem na psicologia e neurobiologia, para imitar os neurônios, sendo um conjunto de entradas e saídas conectadas no qual cada conexão tem um peso associado. Durante a fase de aprendizado, a rede aprende ajustando os pesos para ser capaz de predizer a correta classificação. De acordo com Kirk (2014), o algoritmo de redes neurais tem desempenho superior em função de aproximação e em aprendizado supervisionado; porém, é limitado em problemas com entrada binária. As redes neurais necessitam de parâmetros determinados empiricamente, de acordo com Han e Kamber (2006), tal como a topologia da rede ou estrutura, envolvendo um treinamento longo e, portanto, adequadas a diversas situações, como reconhecimento de caracteres escritos a mão, em patologias e para treinar um computador para pronunciar texto em inglês. As vantagens de utilização é a tolerância a ruído nos dados e a habilidade de classificar padrões os quais não foram empregados na etapa de treinamento.
- Regras de Associação: de acordo com Hall, Witten e Frank (2011), regras de associação são como regras de classificação, que para encontrá-las deve-se executar os procedimentos de indução de regra para cada possível combinação de atributos. A quantidade de regras pode ser reduzida através da fixação de medidas como alavancagem, suporte, qui-quadrado, coverage, entre outras. A partir destas medidas, é possível selecionar as regras interessantes para o estudo.
- Árvores de Decisão: segundo Zumel, Mount e Porzak (2014), são úteis quando as variáveis de entrada interagem com as de saída, ou quando as variáveis de entrada relacionam-se entre si, ou são redundantes. Uma árvore de decisão, de acordo com Han e Kamber (2006),

é um fluxograma como estrutura de árvore, onde cada nó interno representa um teste em um atributo e cada ramo, um resultado do teste, e cada nó folha (ou nó terminal) contém um rótulo de classe. Segundo Zaki e Jr (2006), uma árvore de decisão usa um eixo paralelo hiperplano para dividir os dados em R em regiões. Cada uma destas regiões é dividida recursivamente até que os pontos estejam relativamente homogêneos em suas classes.

- *Text Mining*: segundo Hall, Witten e Frank (2011), é a procura por padrões em texto, análogo à mineração de dados, ou seja, o processo de extrair informação útil. Porém, textos são não estruturados, sem forma, e difícil de lidar, mas é o meio mais comum de troca de informação formal na cultura ocidental. Quanto aos campos de utilização de *Text Mining*, Liu, Chen e Ho (2015) citam os seguintes:
  - novas categorias de texto; em documentos de patentes desenvolvendo métodos de recuperação de acordo com semelhanças de texto;
  - na determinação de emails spams pela análise de seu texto;
  - em comentários para descobrir os termos principais;
  - as atitudes e sentimentos dos autores;
  - em textos científicos, para projetar pesquisa inteligente para comparar semelhanças entre textos.

As seções seguintes apresentam o referencial sobre máquina de vetor suporte, redes neurais artificiais e análise de sobrevivência.

3.3 Máquina de vetor suporte

As máquinas de vetor suporte (SVM) têm sido aplicadas a diferentes áreas de conhecimento, de acordo com Han, Kamber e Pei (2011), tanto para a classificação como para regressão. Inicialmente, a descrição de SVM é feita para classificação e em seguida, para regressão.

3.3.1 Máquinas de vetor suporte para classificação

Os dados de treinamento consistem de N pares $(x_1,y_1),(x_2,y_2),...,(x_N,y_N)$, com $x_i \in R^p$ e $y_i \in [-1,1]$. Segundo Hastie, Tibshirani e Friedman (2009), um hiperplano é definido como:

$$x : f(x) = x^T\beta + \beta_0 = 0 \tag{4}$$

onde β é um vetor unitário $\|\beta\| = 1$.

Considerando duas classes, conforme Leskovec, Rajaraman e Ullman (2014), uma máquina de vetor suporte seleciona um hiperplano particular que, além de separar os pontos em duas classes, maximiza a margem, que é a distância entre o hiperplano e os pontos mais próximos do conjunto de treinamento.

Segundo Zaki e Jr (2006) cada uma destas duas partes possui pontos apenas em uma classe, o conjunto de dados é dito linearmente separável e conforme Hastie, Tibshirani e Friedman (2009), pode ser formulado como:

$$\underset{\beta,\beta_0}{Min\|\beta\|} \tag{5}$$

sujeito a $y_i(x_i^T\beta+\beta_0)\geq 1, i=1,...,N$, desconsiderando a restrição de β.

Máquinas de vetor suporte podem lidar com pontos não separáveis com a introdução de variáveis de folga.

Hastie, Tibshirani e Friedman (2009) definem as variáveis de folga, $\xi=(\xi_1,...,\xi_N)$, como o valor de ξ na restrição $y_i(x_i^T\beta+\beta_0)\geq M(1-\xi_i)$, ou a quantidade proporcional pela qual a predição $f(x_i)=x_i^T\beta+\beta_0$ está no lado errado da margem.

Ao delimitar o valor de soma $\sum\xi_i$, podem ocorrer classificações erradas, quando $\xi_i\geq 1$, com o número total destas classificações equivocadas igual a K. Então, pode-se definir $M=1/\|\beta\|$ e reescrever na forma:

$$\min\|\beta\| \tag{6}$$

$$\text{sujeito a} \begin{cases} y_i(x_i^T\beta+\beta_0)\geq 1-\xi_i \quad \forall i \\ \xi_i\geq 0 \\ \sum\xi_i \leq \text{ constante} \end{cases}$$

Conforme Hastie, Tibshirani e Friedman (2009), esta é a forma usual que o classificador de vetor de suporte é definido para o caso não separável. Segundo Zaki e Jr (2006), neste caso, também chamado de margens soft, a meta de classificação é encontrar o hiperplano com máxima margem, que minimiza o termo de folga, tratando-se de um problema de otimização convexo que pode ser resolvido por multiplicadores de Lagrange, e ser representado como:

$$Min_{\beta,\beta_0}\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^{N}\xi_i \tag{7}$$

$$\text{sujeito a} \begin{cases} \sum\xi_i\geq 0, \\ y_i(x_i^T\beta+\beta_0)\geq 1-\xi_i, \forall i \end{cases}$$

onde C é o parâmetro de penalização ou custo. Escolhe-se C grande quando o objetivo não é classificar erroneamente pontos, mas aceitar uma margem estreita. Escolhe-se C pequeno quando o objetivo é que a maioria dos pontos localizem-se longe do limite (ou seja, com margem grande), porém, de acordo com Rajaraman et al. (2012), com alguns pontos mal classificados.

Para a extensão para m classes, de acordo com Han, Kamber e Pei (2011), uma abordagem simples é treinar m classificadores, um para cada classe (onde o classificador j retorna um valor positivo para a classe e negativo para as demais.)

Conforme Hastie, Tibshirani e Friedman (2009), ajusta-se um classificador de vetor suporte usando $h(x_i)=(h_1(x_i),h_2(x_i),...,h_M(x_i)), i=1,...,N$, à função base $h(x), m=1,...,M$,

produzindo a função não linear $\hat{f}(x) = h(x)^T\hat{\beta} + \hat{\beta}_0$. O classificador de máquina de vetor suporte é uma extensão dessa idéia, onde a dimensão do espaço ampliado pode ficar muito grande, infinita em alguns casos. A função Lagrangeana primal é escrita como:

$$LP = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi)] - \sum_{i=1}^{N}\mu\xi \tag{8}$$

onde $\beta = \sum_{i=1}^{N}\alpha_i y_i x_i$,
$0 = \sum_{i=1}^{N}\alpha_i y_i$,
$\alpha_i = C - \mu_i$ e
$\alpha_i, \mu_i, \xi \geq 0, \forall i$

A função Lagrangeana dual tem a forma:

$$LD = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}(\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_i y_i'\langle h(x_i)h(x_{i'})\rangle) \tag{9}$$

Conforme Hastie, Tibshirani e Friedman (2009), a solução pode ser escrita como:

$$f(x) = h(x)^T\beta + \beta_0 \tag{10}$$
$$f(x) = \sum_{i}^{N}\alpha_i y_i\langle h(x), h(x_i)\rangle + \beta_0$$

ado α_i, β_0 pode ser determinado pela solução de $y_i f(x_i) = 1$ para $0 < \alpha_i < C$. As Máquinas de Vetor Suporte utilizam um mapeamento implícito dos dados de entrada definido por uma função de kernel, isto é, uma função retornando o produto interno entre as imagens de dois pontos de dados x, x' no espaço de características. Conforme Hornik, Meyer e Karatzoglou (2006), se uma projeção $\phi X \to H$ é usada, o produto escalar $\langle\phi(x)\phi(x')\rangle$ pode ser representado por uma função kernel:

$$k(x,x') = \langle\phi(x)\phi(x')\rangle \tag{11}$$

Computando o produto interno no espaço transformado, K deve ser um função simétrica (semi) definida positiva. Segundo Hastie, Tibshirani e Friedman (2009), as três escolhas para K podem ser:

Polinomial de grau d: $K(x,x') = (1 + \langle x,x'\rangle)^d$,

Radial: $K(x,x') = \exp(-\gamma\|x-x'\|^2)$,

Rede neural: $K(x,x') = \tanh(\kappa_1\langle X,X'\rangle + \kappa_2)$

Então, conforme Hastie, Tibshirani e Friedman (2009), a solução pode ser escrita como:

$$min_{\beta\beta_0}[\sum_{i=1}^{N}1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2 \tag{12}$$

Com $f(x) = h(x)^T\beta + \beta_0$, "+" indicando a parte positiva, com o objetivo de minimizar $EL(Y, f(X))$. Se $\lambda = 1/C$, então, a solução é igual a $Min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$.

Segundo Hastie, Tibshirani e Friedman (2009), se os dados são separáveis, então, o limite de $\hat{\beta_\lambda}$ e $\lambda \to 0$ definem o hiperplano de separação ótima. Sendo K:

$$K(x, x') = \sum_{(m=1)}^{\infty} \Phi_m(x)\Phi_m(x')\delta_m \tag{13}$$

Com $h_m(x) = \sqrt{\delta_m}\Phi_m(x)$ e $\theta_m = \sqrt{\delta_m}\beta_m$. h_m pode ser descrito em termos de estimativa de função na reprodução espaços do kernel Hilbert, supondo que a base h se origine da auto-expansão, possivelmente finita, de um kernel definido positivo K. Maiores detalhes podem ser consultados em Hastie, Tibshirani e Friedman (2009). Desta forma, é possível escrever:

$$Min_{\beta\beta_0} \sum_{i=1}^{N} (1 - y_i f(x_i))_+ + \frac{\lambda}{2}\alpha^T K\alpha \tag{14}$$

onde K é a matriz de avaliações de kernel para todos os pares de treinamento. Eles podem ser escritos como

$$Min_{f\in H}[1 - y_i f(x_i)]_+ + \lambda J(f) \tag{15}$$

onde H é o espaço estruturado de funções e $J(f)$ um regularizador apropriado nesse espaço. Se o número de características for grande, de acordo com Hsu et al. (2003), o mapeamento não linear não melhora o desempenho e o uso da função Kernel linear é boa o suficiente, com apenas uma busca para o valor de C.

Segundo Hsu et al. (2003), a realização de uma análise utilizando máquina de vetor suporte seguem as etapas:

1. Transformar os dados. Os dados categóricos necessitam ser colocados no formato numérico e, para tanto, os autores sugerem usar m variáveis para um atributo com m categorias. Para cada uma destas novas variáveis, para a categoria de interesse, o valor é 1, para os demais, 0;
2. Escalonar os dados, no intervalo [-1;1] ou [0;1];
3. Considerar a função Kernel (RBF- radial basis function);
4. Usar Cross-Validation para encontrar os melhores valores de C, custos e γ. Na função Kernel RBF, há dois valores a serem escolhidos. O objetivo é escolher os melhores valores para predizer, com acurácia, no conjunto de testes, os valores de classificação.
5. Usar os melhores valores de C e γ para treinar o conjunto de dados;
6. Testar.

Para Zumel, Mount e Porzak (2014), máquinas de vetor suporte são úteis quando há muitas variáveis de entrada ou quando estas interagem com a variável de saída ou entre elas de forma não linear, tendo poucas suposições sobre as distribuições.

3.3.2 Regressão

A ideia básica, segundo Hall, Witten e Frank (2011), é encontrar uma função que aproxima os pontos de treinamento, minimizando o erro de previsão, como em regressão linear. Porém, há diferenças entre a regressão linear e a regressão por Máquina de Vetor Suporte:

- Em SVM, todos os desvios até um parâmetro especificado pelo usuário são simplesmente descartados;
- Ao minimizar o erro, o risco de *overfitting* é reduzido ao tentar maximizar a planificação da função;
- O erro absoluto das previsões é minimizado, em vez do erro quadrado usado na regressão linear.

SVM foi originalmente desenvolvida para classificação, representando o limite de decisão em termos de um subconjunto pequeno de todos os exemplos de treinamento, os vetores de suporte. Para que essa propriedade fosse transferida para o caso da Regressão, foi criada a chamada ε função de perda insensível.

De acordo com Hastie, Tibshirani e Friedman (2009), para estimar β no modelo de regressão,

$$f(x) = x^T\beta + \beta_0 \tag{16}$$

Minimiza-se :

$$H(\beta, \beta_0) = \sum_{i}^{N} V_\varepsilon(y_i - f(x_i)) + \frac{\lambda}{2}||\beta||^2 \tag{17}$$

Onde :

$$V_\varepsilon(r) = \begin{cases} 0 \text{ se } |r| \leq \varepsilon \\ |r| - \varepsilon, \text{ cc.} \end{cases} \tag{18}$$

$\varepsilon-$insensível é uma medida de erro que, segundo Hastie, Tibshirani e Friedman (2009), ignora erros de tamanho menor que ε. Em regressão, estes pontos com erro pequeno são aqueles com pequenos resíduos. De acordo com Schölkopf et al. (1999), o parâmetro ε pode ser útil se a precisão desejada da aproximação puder ser previamente especificada.

De acordo com Hastie, Tibshirani e Friedman (2009), a solução é dada por:

$$\hat{\beta} = \sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i) x_i \tag{19}$$

$$f(x) = \sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i) \langle x, x_i \rangle + \beta_0$$

onde $\hat{\alpha}_i^*$ e $\hat{\alpha}_i$ são positivos e resolvem a seguinte equação:

$$min_{\alpha_i \alpha_i^*} \varepsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{N} y_i (\alpha_i^* + \alpha_i) + \frac{1}{2} \sum_{i=1, i'=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_{i'}^* - \alpha_{i'}) \tag{20}$$

sujeito a: $0 \leq \alpha_i, \alpha_i^* \leq \frac{1}{\lambda}$,
$\sum_{i=1}^{N} (\alpha_i^* - _i) = 0$,
$\alpha_{i_i}^* = 0$

Normalmente, segundo Hastie, Tibshirani e Friedman (2009), apenas um subconjunto dos valores da solução $(\alpha_{i'}^* - \alpha_{i'})$ é diferente de zero, e os valores de dados associados são chamados de vetores de suporte. ε é um parâmetro da função perda V_ε.

Para Hastie, Tibshirani e Friedman (2009), ambos V_ε e V_H dependem da escala de y e, portanto, de r. Se a resposta for escalonada, então, é possível usar valores pré-definidos para ε e C. Já a quantidade λ pode ser estimada por *cross-validation*.

Por outro lado, nos casos em que se deseja que a estimativa seja a mais precisa possível, sem comprometimento com um certo nível de precisão, pode ser utilizado a $\upsilon-$regressão, proposta por Schölkopf et al. (1999). Para cada ponto $_i$, é permitido um erro ε. A partir de variáveis de folga ξ^* e ε, a solução consiste em minimizar:

$$\min \tau(\xi^{(*)}, \varepsilon) = \frac{1}{2} \|w\| + C(\upsilon\varepsilon + \frac{1}{l} \sum_{i=1}^{l} (\xi_i + \xi_i^*)) \tag{21}$$

sujeito a $((wx_i) + b) - y_i \leq \xi + \xi_i$ e
$y_i - ((wx_i) + b) \leq \xi + \xi_i$
$\xi_i^* \geq 0, \xi > 0$

Conforme Schölkopf et al. (1999), introduzindo um multiplicador Lagrangeano, $\alpha_i^{(*)}, \eta_i^{(*)}, \beta \geq 0$ e substituindo o kernel por:

$$k(x, y) = (\phi(x)\phi(y)) \tag{22}$$

Isto leva ao problema de otimização $\upsilon - SVR$: para $\upsilon > 0, C > 0$, definido como:

$$W(\alpha*) = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i)(_j^* - \alpha_j) k(x_i x_j) \tag{23}$$

sujeito a $\sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0$,
$0 \leq \alpha_i^* \geq \frac{C}{l}$,
$\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) \leq C\upsilon$

A regressão estimada é da forma:

$$f(x) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)k(x_i, x_i) + b \tag{24}$$

$\varepsilon = 0$, se:

- $\upsilon > 1$, desde que não haja perda por aumentar ε.
- $\upsilon \leq 1$ e se os dados são livres de ruído e podem ser perfeitamente interpolados com um modelo de baixa capacidade.

De acordo com Schölkopf et al. (1999), assumindo $\varepsilon > 0$, então υ é um limite superior na fração de erros e um limite inferior na fração de vetores de suporte.

De acordo com Hall, Witten e Frank (2011), o conceito de hiperplano de margem máxima é aplicado somente à classificação, porém, para regressão, o algoritmo de máquina de vetor suporte também produz um modelo que pode ser expresso com poucos vetores de suporte e aplicável a problemas não lineares através de funções de Kernel.

De acordo com Hastie (2016), supondo a aproximação para a função de regressão um conjunto de função base $h_m(x), m = 1, 2, ..., M$:

$$f(x) = \sum_{m=1}^{N} \beta_m h_m(x) + \beta_0 \tag{25}$$

Para estimar β e β_0, minimiza-se

$$H(\beta, \beta_0) = \sum_{i=1}^{N}(y_i - f(x_i)) + \frac{\lambda}{2}\sum \beta_m^2 \tag{26}$$

Para alguma medida geral de erro $V(r)$, a solução para $\hat{f}(x)$ tem a forma:

$$\hat{f}(x) = \sum_{i=1}^{N} \hat{a}_i K(x, x_i) \tag{27}$$

com $K(x, y) = \sum_{m=1}^{M} h_m(x)h_m(y)$, com a mesma forma tanto na expansão da função de base radial quanto na estimativa de regularização.

Seja

- $V(r) = r^2$
- H uma matriz de base NxM com elementos de base $h_m(x_i)$
- $M > N$, grande
- $\beta_0 = 0$

β pode ser minimizado pelo critério de penalização dos mínimos quadrados, de acordo com Hastie, Tibshirani e Friedman (2009):

$$H(\beta) = (y - H\beta)^T - (y - H\beta) + \lambda||\beta||^2 \tag{28}$$

Segundo Hastie, Tibshirani e Friedman (2009), a solução é dada por $\hat{y} = H\hat{\beta}$, com β igual a:

$$-H^T(y - H\hat{\beta}) + \lambda\hat{\beta} = 0 \tag{29}$$

A matriz HH^T consiste de produtos internos entre pares de observações $ii^{'}$, que, de acordo com Hastie, Tibshirani e Friedman (2009), é a avaliação de um produto interno kernel $HH^T_{ii'} = K(x_i x_{i'})$. Os valores preditos de x satisfazem:

$$\hat{f}(x) = h(x)^T\hat{\beta} = \sum_{i=1}^{N} \alpha_i K\hat{(}x, x_i) \tag{30}$$

onde $\hat{\alpha} = (HH^T + \lambda I)^{-1}y$. O produto kernel $K(x_i x_i')$ necessita ser avaliado em todos os $ii^{'}$ pontos de treinamento e nos x pontos para predição.

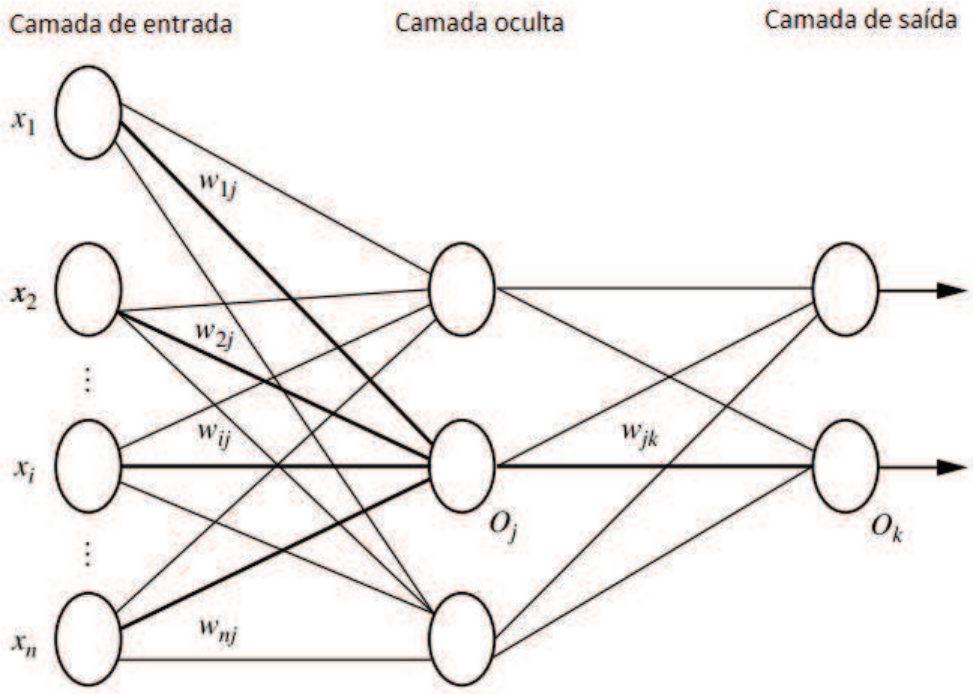
De acordo com Hall, Witten e Frank (2011), a regressão linear padrão tem complexidade $O(m^3)$, à medida que requer a inversão de uma matriz mm, onde m é o número de atributos. A regressão kernel tem complexidade $O(n^3)$, pois envolve uma matriz nn, onde n é o número de instâncias nos dados de treinamento. No entanto, é vantajoso usar a regressão de kernel nos casos em que um ajuste não linear é desejado, ou onde há mais atributos do que instâncias de treinamento.

3.4 Redes Neurais

Uma rede neural artificial (RNA), segundo Lantz (2013), modela a relação entre um conjunto de sinais de entrada e um sinal de saída através do uso de um modelo, tal qual um cérebro biológico responde a estímulos a partir de entradas sensoriais. Isto é, a rede neural artificial tenta resolver problemas de aprendizagem, tal como o cérebro usa uma rede de neurônios. A Figura 11 apresenta uma rede neural através de uma camada de entrada, uma camada oculta e uma de saída.

Para Günther e Fritsch (2010), redes neurais são extensões de modelos lineares generalizados e podem ser aplicados de maneira similar. Segundo o autor, não é necessário especificar o

Figura 11 – Representação de uma rede neural



Fonte: Han, Kamber e Pei (2011)

tipo de relação entre covariáveis e variável resposta, como combinação linear, ao contrário dos modelos lineares generalizados.

Os dados observados são usados para treinar a rede neural e esta aprende por uma aproximação da relação, adaptando iterativamente seus parâmetros. A camada de entrada consiste das covariáveis em neurônios separados, e a camada de saída, da variável resposta.

Embora, segundo Lantz (2013), haja variantes, redes neurais podem ser definidas em termos de: i) uma função de ativação, a qual transforma a combinação de neurônios de entrada, em um sinal de saída a ser transmitido na rede; ii) uma arquitetura de rede, descrevendo o número de neurônios e de camadas e a forma em que elas estão conectadas e iii) o algoritmo de treinamento, que especifica como os pesos de conexão são definidos em proporção ao sinal de entrada.

Hastie, Tibshirani e Friedman (2009) definem rede neural como um modelo de regressão ou classificação de dois estágios, tipicamente representado por um diagrama de rede, aplicável tanto à regressão quanto à classificação. Para regressão, normalmente $K = 1$ e há apenas uma unidade de saída $Y1$.

Os parâmetros de redes neurais são melhor determinados de forma empírica, como a topologia de rede ou estrutura. De acordo com Hastie, Tibshirani e Friedman (2009), os recursos Z_m são criados a partir de combinações lineares das entradas, e a função das combinações lineares

de Z_m modela o alvo Y_k, na forma:

$$\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, ..., M \\ T_k &= \beta_{0k} + \beta_k^T Z, k = 1, ..., K \\ f_x(K) &= g_k(T), k = 1, ..., K \end{aligned} \tag{31}$$

onde $Z = (Z_1, Z_2, ..., Z_M)$ e $T = (T_1, T_2, ..., T_K)$. $\sigma(v)$ é a função de ativação. O design ainda pode conter uma unidade adicional de vício, alimentando cada unidade nas camadas oculta e de saída. A função $g_k(T)$ é a função de saída. Para a regressão, normalmente escolhe-se a função de identidade $g_k(T) = T_k$.

Conforme Hastie, Tibshirani e Friedman (2009), o conjunto completo de pesos é denotado por θ, consistindo de: $\alpha_{0m}, \alpha_m, m = 1, 2, ..., M$, $M(p+1)$ e $\beta_{0k}, \beta_k, k = 1, 2, ...K$, $K(M+1)$

Conforme Han, Kamber e Pei (2011), considerando que cada conexão tem um peso (α ou β), denotados por w na equação 32, para computar a entrada I à unidade j, cada saída O é multiplicada pelo peso correspondente e então, somados. No caso de uma camada oculta, a conexão tem a forma:

$$I_j = \sum_o w_i j O_i + vicio \tag{32}$$

Onde o vício age como um limite, pois serve para variar a atividade da unidade.

De acordo com Günther e Fritsch (2010), a função de erro pode ser de duas formas: ou como o erro médio quadrático ou entropia cruzada. Para regressão, é utilizada o erro quadrático:

$$R\theta = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2 \tag{33}$$

Com as derivadas:

$$\begin{aligned} \frac{\delta R_i}{\delta \beta_{km}} &= -2(y_i k - f_k(x_i)) g' k(\beta_k^T z_i) z_{mi} \\ \frac{\delta R_i}{\delta \alpha_{ml}} &= -\sum_{K}^{k=1} 2(y_i k - f_k(x_i)) g' k(\beta_k^T z_i) \beta_{km} \sigma'(\alpha_m^T x_i) x_{il} \end{aligned} \tag{34}$$

E estas podem ser reescritas como:

$$\begin{aligned} \frac{\delta R_i}{\delta \beta_{km}} &= \delta_{ki} z_{mi} \\ \frac{\delta R_i}{\delta \alpha_{ml}} &= s_{mi} x_{il} \end{aligned} \tag{35}$$

onde δ_{ki} e s_{mi} são "erros" do modelo atual nas unidades de saída e camada oculta, respectivamente, que satisfazem a equação conhecida como equação de retropropagação:

$$s_{mi} = \sigma'(\alpha_m^T x_i) \sum_{k=1}^{K} \beta_{km} \delta_{ki} \tag{36}$$

De acordo com Hastie, Tibshirani e Friedman (2009), utilizando esta equação são obtidas as seguintes atualizações:

$$\begin{aligned} \beta_{km}^{r+1} &= \beta_{km}^{r} - \gamma_r \sum_{i=1}^{N} \frac{\delta R_i}{\delta \beta_{km}^{r}} \\ \alpha_{ml}^{r+1} &= \alpha_{ml}^{r} - \gamma_r \sum_{i=1}^{N} \frac{\delta R_i}{\delta \alpha_{ml}^{r}} \end{aligned} \tag{37}$$

Geralmente, de acordo com Hastie, Tibshirani e Friedman (2009), as redes neurais têm muitos pesos e superpõem os dados no mínimo global de R. Quanto ao critério de parada, treina-se o modelo por um tempo e a parada é realizada antes do mínimo global. Como os pesos começam com uma solução regularizada (linear), isso tem o efeito de encolher o modelo final em direção a um modelo linear. Um conjunto de dados de validação é útil para determinar quando parar, pois espera-se que o erro de validação comece a aumentar.

Um método para regularização, conforme Hastie, Tibshirani e Friedman (2009), é o decaimento de peso, no qual adiciona-se uma penalidade à função de erro: $R(\theta) + \lambda J(\theta)$ onde:

$$J(\theta) = \sum_{km} \beta_{km}^2 + \sum_{lm} \alpha_{lm}^2 \tag{38}$$

Com $\lambda \geq 0$. Os algoritmos de redes neurais são inerentemente paralelos, podendo ser usados usado processamento paralelo para acelerar o processo de computação. Além disso, várias técnicas foram desenvolvidas para a extração de regras de redes neurais treinadas, o que contribui para a utilidade na classificação e previsão numérica na mineração de dados. (HAN; KAMBER; PEI, 2011).

O mais popular algoritmo é o retropropagação. Este algoritmo executa a aprendizagem em uma rede neural multicamada *feedforward* ou acíclica. Segundo Lantz (2013), as redes neurais em que o sinal de saída é alimentado continuamente até que alcance a camada de saída é chamada rede neural *feedforward* (acíclica).

O aprendizado por retropropagação é um processo no qual compara-se a predição de cada tupla com o valor alvo. Os pesos são então modificados para minimizar o erro entre a predição e o valor observado na direção *backwards* (a partir da camada de saída), através de cada camada oculta até a primeira camada oculta. Em geral, segundo Hall, Witten e Frank (2011), os

pesos convergem e o processo é finalizado. Dada uma estrutura de rede, com camadas ocultas, há o algoritmo gradient descent, que determina pesos apropriados para as conexões na rede.

Por causa da forma composicional do modelo, o gradiente pode ser facilmente derivado usando a regra da cadeia para diferenciação. Isso pode ser calculado por uma varredura direta e reversa pela rede, mantendo o controle apenas das quantidades locais de cada unidade. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Segundo Hastie, Tibshirani e Friedman (2009), as vantagens da retropropagação são sua natureza simples e local. No algoritmo de propagação de retorno, cada unidade oculta passa e recebe informações apenas e para as unidades que compartilham uma conexão. Por isso, pode ser implementado de forma eficiente em um computador de arquitetura paralela.

De acordo com Lantz (2013), uma rede neural com múltiplas camadas ocultas é chamada de *Deep Neural Network* (DNN) e a prática de treinamento é às vezes referida como *deep learning.* A Figura 11 apresentou uma rede neural com uma camada oculta.

Quanto à arquitetura, uma rede neural multi camada consiste em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Conforme Han, Kamber e Pei (2011), cada uma destas camadas é composta de unidades: as entradas correspondem aos atributos, que passam pela camada e são então ponderadas e alimentadas simultaneamente a uma segunda camada de unidades de neurônios, a camada oculta.

As saídas destas unidades da camada oculta podem ser inseridas em outras camadas ocultas, e assim por diante, sendo o número de camadas ocultas arbitrário, embora, na prática, geralmente apenas uma seja usada.

Segundo Lantz (2013), a decisão do número de neurônios ocultos é uma decisão do usuário, anterior ao treinamento do modelo. Da mesma forma não há regras confiáveis para o número de neurônios nas camadas ocultas e dependem do número de nodos de entradas, a quantidade de dados de treinamento, a quantidade de ruído e a complexidade da tarefa de aprendizado, entre outros fatores. Quando a acurácia de uma rede treinada for considerada insatisfatória, repete-se o processo de treinamento com uma topologia ou pesos iniciais diferentes.

De acordo com Hastie, Tibshirani e Friedman (2009), geralmente, é melhor ter muitas unidades ocultas a poucas. Normalmente, o número de unidades ocultas está no intervalo de 5 a 100, com o número aumentando com o número de entradas e o número de casos de treinamento. Para Lantz (2013), um grande número de neurônios resultará em um modelo próximo aos dados de treinamento, mas com o risco de *overfitting*. Por outro lado, o treinamento pode ser computacionalmente caro e lento. A melhor prática, então, é usar menos nodos que resultarão em uma performance adequada.

Conforme Han, Kamber e Pei (2011), é aplicada uma função de ativação não linear às saídas ponderadas da última camada oculta e, então, são inseridas como camadas de entrada nas unidades que compõem a camada de saída, a qual emite a previsão da rede para as tuplas

fornecidas. Segundo Lantz (2013), a mais comumente função de ativação usada é a sigmoide, mais especificamente, a sigmoide logística, a qual produz uma saída entre os valores 0 e 1. Segundo Lewis (2016) a função sigmoide é popular em parte porque pode ser facilmente diferenciado, reduzindo o custo computacional durante o treinamento e é representada por:

$$\sigma(v) = \frac{1}{1+\exp(-cv)} \tag{39}$$

onde c é uma constante. De acordo com Lantz (2013), a escolha da função de ativação conduz a rede neural de forma que possa ajustar certos tipos de dados de maneira mais apropriada, permitindo a construção de redes neurais especializadas. Como exemplo, os autores citam que a função de ativação linear resulta em uma rede neural similar ao modelo de regressão linear, enquanto a função gaussiana resulta em uma rede neural chamada função de base radial.

As variáveis de entrada na fase de treinamento devem ser normalizadas, de acordo com Han, Kamber e Pei (2011), geralmente em valores entre 0 e 1 e as variáveis com valores discretos com k categorias são transformados em k variáveis assumindo valores 0 e 1.

A taxa de aprendizagem γ_r é geralmente considerada uma constante e pode ser otimizada por uma pesquisa que minimiza a função de erro em cada atualização. Com o aprendizado *online*, γ_r deve ir para zero com a iteração $r \to \infty$. Os resultados asseguram a convergência se $\gamma_r \to 0$, $\sum_r \gamma_r = \infty$ e $\sum_r \gamma_r^2 < \infty$ (satisfeito, por exemplo, por $\gamma_r = 1/r$). Porém, segundo Günther e Fritsch (2010), podem ocorrer dificuldades de convergência usando um grande número de covariáveis ou variáveis de resposta.

O algoritmo para de acordo com algum critério pré-especificado. Segundo Günther e Fritsch (2010), a complexidade da função calculada aumenta com a adição de camadas ocultas ou neurônios ocultos. Como critérios de parada podem ser citados o número de iteração.

3.5 Análise de Sobrevivência

Análise de sobrevivência é caracterizada por apresentar duas variáveis consistindo como resposta: o tempo até a ocorrência de um evento de interesse, também chamado de tempo de falha, e a presença de censura ou observação parcial da resposta.

Em relação à censura, de acordo com Liu (2012), como os dados são geralmente coletados ou por um intervalo de tempo em que a ocorrência de um evento particular é observado, ou como janela entre dois tempos limites, pode acontecer que tempos completos de sobrevivência, para muitas das unidades, não sejam observados, com perda de informação, o que se denomina de dados censurados. De acordo com Johnson (1999), a censura é uma variável binária, podendo ser medida através de experimento, contagem direta ou por estimação indireta. Segundo Colosimo e Giolo (2006), sem a presença das censuras, as técnicas estatísticas clássicas como regressão

e planejamento de experimentos podem ser usadas, provavelmente, com a transformação da variável resposta.

A censura pode ocorrer de várias formas ou razões, de acordo com Allison (2010), podendo ser à direita ou à esquerda. Censura à direita ocorre quando a observação é encerrada antes que o evento de interesse ocorra, enquanto a censura à esquerda ocorre quando se começa a observar uma variável, porém, o evento de interesse já ocorreu. Outro conceito é o truncamento. De acordo com Colosimo e Giolo (2006), o truncamento é caracterizado por uma condição, assim, os valores perdidos (*missing values*) não são censura, mas dados truncados.

De acordo com Liu (2012), um considerável corpo de análise de sobreviência é conduzida através de regressão com dados censurados envolvendo uma ou mais variáveis preditoras. Com a adição de covariáveis, a análise de sobrevivência pode ser vista como consistindo de informação sobre três fatores primários: tempo de sobrevivência, estado censurado e covariáveis.

Segundo Allison (2010), há vários métodos para a análise de sobrevivência: tabelas de vida, estimador de Kaplan Meyer, regressão exponencial, regresão lognormal, entre outros, sendo, muitas vezes, complementares.

3.5.1 Funções características

De acordo com Klein e Moeschberger (2005), seja X o tempo até a ocorrência de um evento específico. Quatro funções caracterizam a distribuição de X:

- função de sobrevivência, ou a probabilidade de sobrevivência ao tempo x, no contexto da manufatura, é referida como a função de confiabilidade:

$$S(x) = Pr(X > x) \tag{40}$$

  Se X é uma variável aleatória contínua, então, $S(x) = 1 - F(x)$, onde $F(x)$ é a função densidade de probabilidade. Se for uma variável aleatória discreta, $S(x) = \sum_{x_j > x} p(x_j)$

- função de risco: também conhecida como taxa de falha condicional em confiabilidade:

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x} \tag{41}$$

  onde:

  - $h(x) = -dln[S(x)]/dx$, se x é uma v.a. contínua

  - $h(x_j) = p(x_j)/S(x_{j-1}), j = 1, 2, ..., S(x_0) = 1$ se x é uma v.a discreta.

  A função de risco geralmente é mais informativa sobre o mecanismo de falha subjacente do que a função de sobrevivência, o que faz com que esta função seja o método dominante para resumir os dados de sobrevivência.

- função massa de probabilidade: probabilidade incondicional de ocorrência do evento no tempo x;
- a vida residual média no tempo x, ou o tempo médio para o evento de interesse, dado que o evento não ocorreu em x:

$$mrl(x) = E[X - x|X > x] \tag{42}$$

Em análise de sobrevivência, a modelagem pode ser feita a partir de modelos paramétricos, não paramétricos e semi paramétricos. Na estimação não paramétrica, a função de sobrevivência pode ser pelo estimador de Kaplan Meyer, de Nelson-Aalen ou ainda por tabela de vida atuarial. O modelo de regressão de Cox, também conhecido por modelo semi paramétrico, no qual é possível incorporar covariáveis, ao contrário dos modelos não-paramétricos. Os Modelos regressão paramétricos permitem a inclusão de covariáveis, são mais eficientes, porém menos flexíveis do que os semi paramétricos.

A Tabela 1 traz a taxa de risco, função de sobrevivência, função densidade de probabilidade e tempos de vida esperados para algumas distribuições paramétricas.

Tabela 1 – Taxa de risco, função de sobrevivência, função de densidade de probabilidade e tempos de vida esperados

Distribuição	Função de risco	Função de Sobrevivência	Função densidade de probabilidade
Exponencial $\lambda > 0, x \geq 0$	λ	$exp(\lambda x)$	$\lambda exp(-\lambda x)$
Weibull $\alpha, \lambda > 0, x \geq 0$	$\alpha\lambda x^{\alpha-1}$	$exp(-\lambda x^{\alpha})$	$\alpha\lambda x^{\alpha-1} exp(-\lambda x^{\alpha})$
Gamma $\beta, \lambda > 0,$ $x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - I(\lambda x, \beta)^{*}$	$\frac{\lambda^{\beta} x^{\beta-1} exp(-\lambda x)}{\Gamma(\beta)}$
Log normal $\sigma > 0, x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - \phi[\frac{lnx-\mu}{\sigma}]$	$\frac{exp[-\frac{(lnx-\mu)^2}{2\sigma^2}]}{x(2\pi)^{1/2\sigma}}$
Log logística $\alpha, \lambda < 0, x \leq 0$	$\frac{\alpha x^{\alpha-1}\lambda}{1+\lambda x^{\alpha}}$	$\frac{1}{1+\lambda x^{\alpha}}$	$\frac{\alpha x^{\alpha-1}\lambda}{[1+\lambda x^{\alpha}]^2}$
Normal $\sigma > 0,$ $-\infty < x, \infty$	$\frac{f(x)}{S(x)}$	$1 - \phi[\frac{x-\mu}{\sigma}]$	$\frac{exp[-\frac{(x-\mu)^2}{2\sigma^2}]}{(2\pi)^{1/2\sigma}}$
Exponencial power $\alpha, \lambda > 0, x \geq 0$	$\alpha\lambda^{\alpha} x^{\alpha-1} exp[(\lambda x)^{\alpha}]$	$exp[1 - exp[(\lambda x)^{\alpha}]$	$\alpha e \lambda^{\alpha} x^{\alpha-1} exp[(\lambda x)^{\alpha}]$ $-exp[exp[(\lambda x^{\alpha})]]$
Gompert $\Theta, \alpha > 0, x \geq 0$	$\Theta e^{\alpha x}$	$exp[\frac{\Theta}{\alpha}(1 - e^{\alpha x})]$	$\Theta e^{\alpha x} exp[\frac{\Theta}{\alpha}(1 - e^{\alpha x})]$
Pareto $\Theta > 0, \lambda > 0$ $x \geq 0$	$\frac{\Theta}{x}$	$\frac{\lambda^{\Theta}}{x^{\Theta}}$	$\frac{\Theta\lambda^{\Theta}}{x^{(\Theta+1)}}$
Gamma Generalizada $\lambda > 0, \alpha > 0$ $\beta > 0, x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - I[\lambda x^{\alpha}, \beta]$	$\frac{\alpha\lambda^{\beta} x^{\alpha\beta-1} exp(-\lambda x^{\alpha})}{\Gamma(\beta)}$

Fonte: Klein e Moeschberger (2005)

3.5.2 Estimação não paramétrica

Seja Y_i o número de indivíduos que estão sob o risco no tempo i e d_i é o número de mortes no tempo i . Segundo Klein e Moeschberger (2005), considerando dados censurados à direita, a função de sobrevivência estimada proposta por Kaplan e Meyer, conhecida também como estimador do produto limite é:

$$\hat{S}(t) = 1, \quad \text{Se t} < \text{t}_1,$$

$$\prod_{t_i<t}[1-\frac{d_i}{Y_i}] \quad \text{se } t_1 < t \tag{43}$$

E a variância é dada pela fórmula de Greenwood:

$$\hat{V}\hat{S}(t) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)} \tag{44}$$

A estimação da função de risco acumulada é dada por:

$$\hat{H}(t) = -ln[\hat{S}(t)] \tag{45}$$

Uma alternativa ao estimador de Kaplan Meyer, para a função de risco acumulada, de acordo com Klein e Moeschberger (2005), é referido como o estimador de Nelson–Aalen, dado por:

$$\hat{H}(t) = 0, \text{ se } t < t_1, \tag{46}$$

$$\sum_{t_i \leq t} \frac{d_i}{Y_i}, \text{ se } t_i \leq t$$

E a variância do estimador é:

$$\sigma^2_H(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i^2} \tag{47}$$

Com base no estimador de Nelson Aalen, segundo Klein e Moeschberger (2005), a função de sobrevivência é dada por:

$$\tilde{S}(t) = \exp[-\tilde{H}(t)] \tag{48}$$

O intervalo de confiança, para um único tempo fixo no qual a inferência é feita, é dado por

$$\hat{S}(t_0) \pm Z_{1-\alpha/2}\sigma_S(t_0)\hat{S}(t_0) \tag{49}$$

onde $\sigma^2_{S(t)} = \frac{\hat{V}\hat{S}(t)}{\hat{S}^2(t)}$ e $Z_{1-\alpha/2}$ é o percentil $1-\alpha/2$ de uma distribuição normal padrão.

3.5.3 Regressão dos riscos proporcionais semi paramétricos com covariáveis fixas

Seja uma amostra de tamanho n, $(T_j, \delta_j, Z(t)), j = 1, ..., n$, onde T_j é o tempo dos j casos (pacientes, peças, etc.), δ_j é o evento indicador para o caso j, assumindo o valor 1 se o evento ocorre e 0, se é censurado; e Z_t é o vetor de covariáveis ou fatores de risco. Assim, segundo Klein e Moeschberger (2005), o risco proporcional no tempo j, considerando o vetor de covariáveis Z, de acordo com o modelo de Cox é:

$$h(t/Z) = h_0(t)c(\beta^t Z) \tag{50}$$

onde $h_0(t)$ é arbitrário, e tratado de forma não paramétrica, $\beta = \beta_1, ..., \beta_p$ é um vetor de parâmetros e $c(\beta^t Z)$ é uma função conhecida. Este modelo, segundo Klein e Moeschberger (2005), é chamado de modelo semi paramétrico porque uma forma paramétrica é assumida apenas para o efeito da covariável. Um modelo comum para $c(\beta^t Z)$ é:

$$c(\beta^t Z) = exp(\eta^t Z) = exp(\sum_{k=1}^{p} \beta_k Z_k) \tag{51}$$

O modelo de Cox também é chamado de modelo de riscos proporcionais porque, dado dois casos com a mesma covariável Z e Z^*, a razão dos riscos é dada por Klein e Moeschberger (2005):

$$\frac{h(t/Z)}{h(t/Z^*)} = exp[\sum_{k=1}^{p} \beta_k (Z_k - Z_k^*)] \tag{52}$$

Por exemplo, em estudos clínicos, se Z_1 indica o efeito de um tratamento, assumindo valor 1 se o paciente recebe o tratamento e 0 se recebe o placebo e, mantendo todas as outras covariáveis com o mesmo valor, então, $h(t/Z)/h(t/Z^*) = exp(\beta_1)$ é o risco de ter o evento se o indivíduo recebeu o tratamento em relação ao risco de ter o evento, caso o indivíduo tenha recebido o placebo.

De acordo com Colosimo e Giolo (2006), a suposição para o uso do modelo de regressão de Cox é a proporcionalidade das taxas de falhas. Há casos em que a suposição de riscos proporcionais é violada para alguma covariável. Nesses casos, é possível estratificar essa variável e empregar o modelo de risco proporcional dentro de cada estrato para as outras covariáveis. A suposição de riscos proporcionais pode ser testada através do teste de razão de verossimilhança ou pelo teste de Wald. Quanto ao modelo aditivo, pode ser encontrado em Klein e Moeschberger (2005).

De acordo com Klein e Moeschberger (2005), há quatro aspectos nos modelos de riscos proporcionais a serem analisados:

- dada uma covariável, observar a melhor forma funcional para explicar a influência da covariável na sobrevida;

- a adequacidade da suposição de riscos proporcionais;
- a acurácia para predizer a sobrevivência com o objetivo de identificar os casos que falham muito antes ou muito depois do previsto no modelo;
- examinar a influência ou alavancagem de cada caso no ajuste do modelo.

A adequação do modelo de Cox pode ser testada através de gráficos de resíduos. Podem ser examinados os seguintes resíduos: Cox-Snell, Martingale, gráficos baseados na função risco acumulada de um modelo estratificado, e deviance.

3.5.4 Regressão Paramétrica

Os modelos paramétricos, de acordo com Klein e Moeschberger (2005), tendem a fornecer estimativas mais precisas e baseadas em menos parâmetros que os modelos não paramétricos ou semi paramétricos.

Considerando o tempo de falha X, $X > 0$, segundo Klein e Moeschberger (2005), e o vetor $Z^t = (Z_1, ..., Z_p)$ de variáveis explanatórias associadas com o tempo de falha X. Z^t pode incluir variáveis quantitativas, qualitativas e/ou dependentes do tempo.

De acordo com Klein e Moeschberger (2005), a abordagem para modelar o efeito das covariaveis é análoga à regressão linear clássica, sendo a variável $Y = ln(x) : Y = \mu + \gamma^t Z + \sigma W$ onde $\gamma^t = (\gamma_1, ..., \gamma_p)$ é o vetor dos coeficientes de regressão e W é a distribuição do erro.

Este modelo é chamado de modelo de tempo de falhas acelerado e a estimação dos coeficientes é geralmente através de mínimos quadrados. O efeito das variáveis explanatórias no tempo é mudada por um fator $exp(\Gamma^t Z)$, dependendo do sinal de $\Gamma^t Z$. A taxa de risco é dada por Klein e Moeschberger (2005):

$$h(x|Z) = h_0[xexp(-\Gamma^t Z)]exp(\Gamma^t Z) \tag{53}$$

Para os modelos paramétricos a seguir, considera-se que tenham uma representação de modelo de tempo de falha acelerado e uma representação do logaritmo do tempo. A função de sobrevivência é representada por, de acordo com Klein e Moeschberger (2005), é dado por:

$$S(x|Z) = S_0[exp(\Theta^t Z)x], \quad \forall x \tag{54}$$

onde $exp(\Theta^t Z)$ é chamado de fator de aceleração. A taxa de risco individual, considerando a covariável Z é dada por:

$$h(x|Z) = exp(\Theta^t Z)h_0[exp(\Theta^t Z)x], \forall x \tag{55}$$

De acordo com Klein e Moeschberger (2005), os testes estatísticos formais para falta de ajuste tendem a ter baixa potência para amostras pequenas ou rejeitar um dado modelo para

amostras grandes. Assim, as verificações de adequação também podem ser feitas através de gráficos. As verificações de ajuste servem como um meio para rejeitar modelos inapropriados, não para provar que um modelo particular está correto. O gráfico básico é feito pela taxa de risco acumulado e o estimador de Nelson-Aalen. A Tabela 2 representa os parâmetros para teste das distribuições lognormal, Weibull e exponencial. Klein e Moeschberger (2005)

Tabela 2 – Parâmetros para as distribuições

Modelo	Taxa de risco acumulado	Gráfico	Resíduo de Cox Snell
Exponencial	λx	$\hat{H}$ versus x	$r_i = \hat{\lambda} t_i exp[\hat{\beta}^t Z_i]$
Weibull	λx^{α}	$ln\hat{H}$ versus lnx	$\hat{\lambda} exp[\hat{\beta}^t Z_i] t_1^{\hat{\alpha}}$
Log normal	$-ln(1-\phi[ln(x)-\mu]/\sigma)$	$\phi^{-1}[1-exp(-\hat{H})]$ versus lnx	$ln[1-\phi(\frac{lnT_j-\hat{\mu}-\hat{\gamma}^t Z_j}{\hat{\sigma}})]$

Fonte: Klein e Moeschberger (2005)

A verificação do ajuste dos modelos pode ser feita através de gráficos. Um qq-plot é traçado para verificar o correto ajuste dos dados e o estimador de Kaplan-Meyer é calculado. O gráfico resultante deve ser uma reta através da origem, se o modelo de tempo de falha acelerado se mantiver. Duas comparações podem ser feitas para verificar visualmente a adequacidade das distribuições: i) as funções de sobrevivêcia de acordo com as distribuições ajustadas versus a sobrevivência ajustada de Kaplan Meyer, segundo Klein e Moeschberger (2005) e ii) as funções linearizadas.

De acordo com Colosimo e Giolo (2006), a linearização da função de sobrevivência tendo como ideia básica a construção de gráficos que sejam aproximadamente lineares, caso a proposta de modelo seja apropriada. A forma de cálculo para os gráficos estão na tabela 2.

A verificação do ajuste também é realizada através dos gráficos dos resíduos de Cox-Snell e deviance. De acordo com Colosimo e Giolo (2006), o objetivo da análise dos resíduos de Cox-Snell é examinar o ajuste global do modelo e, se o modelo for adequado, estes resíduos devem seguir uma distribuição exponencial.

O gráfico dos resíduos de Cox–Snell pode ser útil para verificar o ajuste total final do modelo e é calculado conforme a Tabela 2. Se o modelo está bem ajustado, o gráfico de $\hat{H}_r(r_j)$ versus r_j é uma reta passando pela origem e com inclinação igual a 1.

Para os modelos paramétricos, os desvios de Martingale são dados por:

$$M_j = \delta_j - r_j \tag{56}$$

E os resíduos deviance são dados por:

$$D_j = sign[M_j](-2[M_j + \delta_j ln(\delta_j - M_j)])^{1/2} \tag{57}$$

Se o modelo está correto, então, os resíduos deviance comportam-se como um ruído aleatório. Os gráficos tanto dos desvios de Martingale como o deviance versus o tempo, o número de observações, o fator de aceleração fornecem uma verificação da adequabilidade do modelo. Segundo Klein e Moeschberger (2005), a interpretação destes gráficos é análoga à análise de resíduos da análise de regressão.

Segundo Jenkins (2005), a estimativa de modelos paramétricos é baseada em expressões para funções de sobrevivência e de densidade. Assim, a descrição de riscos proporcionais versus tempo de vida acelerado refere-se à interpretação de estimativas de parâmetros e não às diferenças de estimação dos modelos.

De acordo com Colosimo e Giolo (2006), a partir da exponencial dos coeficientes estimados no modelo é obtida a razão dos tempos medianos de sobrevivência, mantendo as demais variáveis constantes. Desta forma, os coeficientes podem ser interpretados.

3.5.5 Seleção de variáveis

Jr e Lemeshow (1999) apresentam três formas de escolha de covariáeis, considerando o modelo semi paramétrico: seleção intencional, controlada pelo analista, *stepwise* e a seleção dos melhores subconjuntos. Segundo Colosimo e Giolo (2006), há rotinas automáticas para seleção de covariáveis como os métodos *forward, backward* e *stepwise*. A desvantagem indicada nestes métodos é que eles tendem a identificar um conjunto particular de covariáveis em vez de possíveis conjuntos igualmente bons para a explicação da resposta. O autor sugere que covariáveis consideradas pelo pesquisador como relevantes não sejam excluídas do modelo, independente da significância dos testes estatísticos.

O método *stepwise* para a seleção das covariáveis também pode ser aplicado. Porém, quando o conjunto de covariáveis é grande, a geração do modelo pode apresentar problemas computacionais. Junior e Arrais (2009) apontam, como problema computacional, a instabilidade numérica no processo de maximização da verossimillança, levando a estimativas imprecisas em decorrência da multicolinearidade. Uma possível solução é uma seleção mais criteriosa das covariáveis do modelo.

3.5.6 Tábuas de mortalidade

De acordo com Colosimo e Giolo (2006), uma outra forma de estimação não-paramétrica da função de sobrevivência é por meio de tábuas de vida ou atuariais, também conhecidas como tábuas de mortalidade. Conforme Brasil (1985), tábua de mortalidade é o instrumento destinado a medir as probabilidades de vida e de morte, apontando o número de pessoas da mesma idade, vivas a cada ano. Embora haja variações, as tábuas de mortalidade são constituídas por seis colunas:

1. x: coluna de idades, em ordem cronológica;

2. l: quantidade de pessoas vivas em cada idade;
3. d: quantidade de pessoas mortas em cada idade;
4. q: probabilidade de morte em cada idade;
5. p: probabilidade de sobrevivência em cada idade;
6. e: expectativa completa de vida para cada idade.

3.6 Regras de Associação

As regras de associação, conforme North (2012), são uma metodologia de mineração de dados introduzida por Agrawal, Imieliński e Swami (1993), que buscam associações frequentes entre atributos em um conjunto de dados. De acordo com Rajaraman et al. (2012), as regras de associação são uma coleção de regras if-then, no formato, se X, então Y.

Segundo Hall, Witten e Frank (2011), os termos são originários da análise de cesta de compras em que os itens são bens em um carrinho e o interesse está na associação entre os itens da cesta:

- Itens: $I = \{x_1, x_2, ..., x_m\}$ são um conjunto de atributos binários.
- Itemset: um conjunto $X \subseteq I$, e k-itemset é o nome dado a um conjunto de itens de cardinalidade k, de acordo com (ZAKI; JR, 2006).
- Transações: Seja a transação um conjunto de elementos $T = \{t_1, t_2, ..., t_n\}$, uma tupla $\langle t, X \rangle$, com $t \in T$ o identificador exclusivo da transação. De acordo com Zumel, Mount e Porzak (2014), a transação é a unidade de agregação de itens diferentes.
- Regras de associação: Uma expressão onde X e Y são itemsets disjuntos ($X, Y \subseteq T$, e $X \cap Y = \emptyset$), denotada por $X \rightarrow Y$. (ZAKI; JR, 2006). $X \rightarrow Y$ é como *Premissa* $\rightarrow$ *Conclusao*. Segundo North (2012), premissa é chamada de antecedente (ou lado esquerdo) e a conclusão é o consequente (ou lado direito).

Segundo Zaki e Jr (2006) a regra de associação se 'X, então Y' significa que toda vez que o conjunto de itens X estiver em uma transação, espera-se que o resultado seja Y com uma determinada confiança. As definições de confiança e suporte são as seguintes:

- Suporte é um número S, $S = sup(X \rightarrow Y) = |t(XY)| = sup(XY)$. O suporte relativo é definido como a fração das transações onde X e Y ocorrem juntos ou como uma estimativa da probabilidade de X e Y:

$$rsup(X \rightarrow Y) = \frac{sup(XY)}{|D|} \tag{58}$$

onde D é uma relação binária no conjunto de itens. De acordo com Zaki e Jr (2006), o suporte mínimo *minsup* é o valor mínimo para $sup(X)$ ou $rsup(X)$.

- Segundo Zaki e Jr (2006), confiança pode ser definida como a probabilidade condicional de uma transação que conter Y dado que ela contém X:

$$c = conf(X \rightarrow Y) = P(Y/X) = \frac{sup(XY)}{sup(X)} \tag{59}$$

O valor *minconf* representa a confiança mínima aceitável que uma regra extraída do conjunto de dados deve ter.

Para Zumel, Mount e Porzak (2014), o objetivo é encontrar todas as regras interessantes no banco de dados com mínimo suporte e confiança. Além de um valor mínimo de suporte, quando necessário, um valor máximo de suporte pode ser introduzido para selecionar as regras relevantes.

De acordo com Zaki e Jr (2006), se $sup(XY) \geq minsup$, então o itemset XY é frequente, bem como a regra. Se $conf(XY) \geq minconf$, a regra é considerada forte.

Para obter as regras de associação $X \rightarrow Y$ aplicáveis a uma fração razoável, o suporte de X deve ser razoavelmente alto, embora segundo Rajaraman et al. (2012), o quão alto seja uma questão relativa: os autores citam o caso de lojas de material de construção, em que um suporte razoavelmente alto é de cerca de 1%, mas, para que a regra tenha um efeito prático, a confiança deve ser razoavelmente alta, por exemplo, 50%.

De acordo com Zhang et al. (2007), existe um tipo especial de regra de associação, a regra de associação de classe (CAR), em que o lado direito da regra é uma classe alvo, enquanto o lado esquerdo pode conter um ou mais atributos.

Há vários algoritmos para obter regras de associação: o algoritmo de Park, Chen, and Yu (PCY), o algoritmo multiestágio, o multihash, o simple, algoritmo aleatorizado, o SON algorithm, Toivonen's algoritmo, e métodos híbridos, segundo (RAJARAMAN et al., 2012). Zaki e Jr (2006) incluem força bruta, *apriori*, eclat, declat, e FP-growth algoritmos.

O algoritmo *apriori*, um dos mais usados, encontra as regras de associação em duas etapas: primeiro, todos os conjuntos de itens com suporte maior que o limite fixado *minsup* são encontrados. Então, de acordo com Scheffer (2001), todos os conjuntos de itens são classificados em lados esquerdo e direito e a confiança das regras é calculada. Somente as regras com uma confiança maior que o limite de confiança *minconf* são retornadas.

De acordo com Hall, Witten e Frank (2011), normalmente, a mineração de regras de associação é focada em padrões frequentes. Porém, estes padrões podem resultar em conclusões óbvias ou resultados esperados. No entanto, há aplicações em que o objetivo é identificar associações pouco frequentes, o que pode ser interessante para verificar relações, casos raros

ou fundamentais. Assim, as chamadas regras raras podem capturar *outliers* ou regras com um número muito pequenos de casos, mas fundamentais para compreender o problema em questão.

Neste sentido, as regras frequentes e raras apresentam informações diferentes sobre o banco de dados. De acordo com Koh e Ravana (2016), padrões freqüentes podem apenas representar o conhecido e esperado, enquanto padrões raros podem representar associações inesperadas ou desconhecidas. No entanto, os dois tipos de conjuntos de itens podem ser úteis, desde que haja um contexto apropriado e um objetivo adequado. Em particular, conjuntos de itens pouco frequentes são relevantes em problemas nos quais as informações de *outlier* são relevantes. Koh e Ravana (2016) citam doenças raras, falhas no equipamento de telecomunicações e itens de supermercados pouco comprados como exemplos.

A seleção de regras raras pode ser controlada através da medida suporte. Além dos valores mínimos de suporte, um valor máximo pode ser fornecido para classificar as regras consideradas raras. Luna, Romero e Ventura (2014) referem-se a regras com pouco suporte e alta confiança como esporádicas e indicam que estas podem ser divididas em regras de associação perfeitamente rara (PRAR) e regras de associação imperfeitamente rara (IRAR). De acordo com Koh e Rountree (2005), as regras são perfeitamente raras quando os conjuntos de itens que formam a regra consistem em itens que estão abaixo do limite máximo de suporte, enquanto regras imperfeitamente raras são aquelas em que o suporte máximo é aumentado para incluir conjuntos com itens acima do suporte máximo desejado. Na notação matemática, uma regra perfeitamente rara contém:

$$conf(X \rightarrow Y) \geq c \wedge \forall I \in (X \cup Y), rsup(I) \leq s \tag{60}$$

e uma regra imperfeitamente rara:

$$conf(X \rightarrow Y) \geq c \wedge rsup(X \cup Y) < s \wedge \exists I \in (X \cup Y), rsup(I) \geq s \tag{61}$$

De acordo com Luna, Romero e Ventura (2014) e Koh e Rountree (2005), quando o objetivo é identificar regras raras, o suporte mínimo é baixo, o que torna o uso os algoritmos conhecidos, como *a priori*, inapropriados devido ao seu tempo de processamento e complexidade: à medida que o algortimo varre todo o banco de dados, sua eficiência é afetada, principalmente com grande volume de dados.

De acordo com Koh e Rountree (2005), o algoritmo *apriori*-Inverso é usado para gerar regras raras. O algoritmo mantém os itens quando o valor de suporte é maior que um limite mínimo e menor que um valor máximo. Então, de acordo com Luna, Romero e Ventura (2014), com um limite de confiança, um conjunto de regras de associação pode ser obtido a partir de todas as combinações possíveis de itens previamente obtidos.

Além das medidas suporte e confiança, a alavancagem também pode ser usada como uma medida de interesse. Alavancagem é usada para representar a confiabilidade da regra, de

acordo com Zaki e Jr (2006) e pode ser expressa como:

$$Alavancagem(X \rightarrow Y) = \frac{P(XY)}{P(X)P(Y)} = \frac{rsup(XY)}{rsup(X)rsup(Y)} = \frac{conf(X \rightarrow Y)}{rsup(Y)} \tag{62}$$

Conforme Zumel, Mount e Porzak (2014), se o valor da alavancagem for próximo a 1, é alta a chance da regra não ser confiável. Quanto maior o valor, maior a probabilidade do padrão ser real.

Segundo Liu, Hsu e Ma (1999), muitas associações descobertas são redundantes ou apresentam pequenas variações de outras, podendo ser removidas sem perda da qualidade da informação. Isto é, se a regra está contida em outra regra e seu valor de alavancagem é menor ou igual ao valor da outra regra, ela pode ser removida.

4 MÉTODO

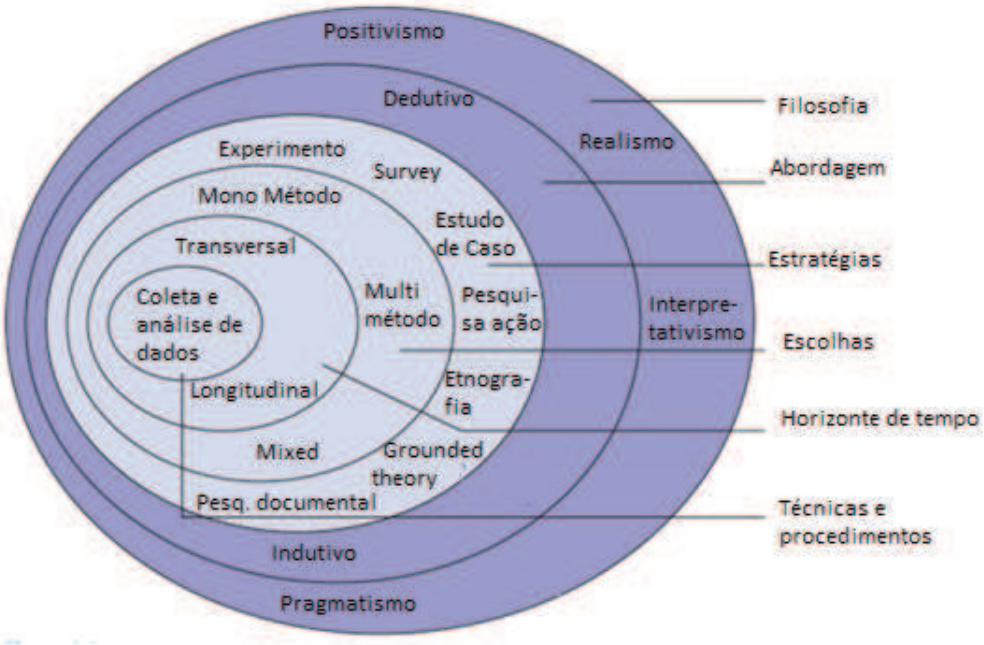
De acordo com Silva e Menezes (2005), pesquisa é um conjunto de ações propostas com o objetivo principal de encontrar a solução para um problema específico, baseada em procedimentos racionais e sistemáticos. Já para Saunders, Lewis e Thornhill (2009), a pesquisa tem o propósito de produzir conhecimento de forma sistemática, no sentido de que é baseada em relações lógicas e não em simples crença, apresentando três características principais:

- os dados são coletados sistematicamente;
- a interpretação é sistemática; e
- o propósito é a descoberta de conhecimento.

4.1 Delineamento da Pesquisa

Saunders, Lewis e Thornhill (2009) consideram uma pesquisa como uma cebola: no centro, estão a coleta e análise de dados, mas para chegar a este ponto, há camadas importantes que necessitam ser definidas, tal como uma cebola. A Figura 12 mostra o modelo cebola para esta pesquisa.

Figura 12 – Modelo Cebola



Fonte: Saunders, Lewis e Thornhill (2009)

O modelo classifica a pesquisa de acordo com a filosofia (positivismo, pragmatismo, interpretativismo e realismo); a abordagem (indutiva, dedutiva); as estratégias (*survey*, estudo de caso, pesquisa ação, experimentos, etnografia, *Grounded Theory*); as escolhas (multi método, monométodo ou método combinado); o horizonte de tempo (transversal ou longitudinal); a coleta e análise de dados. O Quadro 4 classifica esta pesquisa de acordo com as diferentes perspectivas.

Quadro 4 – Classificação da pesquisa conforme o modelo proposto por Saunders, Lewis e Thornhill (2009)

Perspectiva	Classificação
Filosofia	Pós-Positivista
Abordagem	Dedutiva
Estratégias	Modelagem
Escolhas	Monoquantitativo
Horizonte de Tempo	Longitudinal retrospectivo
Técnicas e Procedimentos	Mineração de dados

Fonte: A Autora

O termo filosofia de pesquisa, de acordo com Saunders, Lewis e Thornhill (2009), está relacionado ao desenvolvimento do conhecimento e de sua natureza, isto é, a filosofia de pesquisa adotada contém suposições importantes sobre a forma como o pesquisador vê o mundo e estas suposições irão sustentar as estratégias e os métodos de pesquisa escolhidos.

De acordo com a perspectiva da filosofia, esta pesquisa pode ser classificada como pós positivista à medida que utiliza modelos experimentais com teste de hipóteses para formular teorias explicativas de relações causais.

A classificação, conforme a abordagem, é de uma pesquisa dedutiva, à medida que pretende testar um modelo de predição para o tempo processual. Segundo Saunders, Lewis e Thornhill (2009), a abordagem dedutiva envolve o teste de uma proposição teórica pelo emprego de uma estratégia de pesquisa planejada para o propósito de seus testes, indo do geral para o específico.

A estratégia de pesquisa pode ser utilizada por três propósitos, conforme Yin (2004): descritivo, explanatório ou exploratório. Foi proposto um estudo explanatório com o objetivo de estudar as relações entre as variáveis sendo que, para Triviños (1987), um estudo explanatório estabelece as causas dos fenômenos, determinando quais variáveis atuam e sua influência sobre as demais e para Saunders, Lewis e Thornhill (2009), a ênfase está na explicação entre as variáveis.

Com o objetivo de realizar um estudo explanatório, propõe-se, como estratégia de pesquisa, a Modelagem. Segundo Cauchick e Fleury (2012), a utilização de modelos permite melhor compreender o ambiente em questão, identificar problemas, formular estratégias e oportunidades para apoiar e sistematizar a tomada de decisão. O modelo a ser proposto, quantitativo, compreende um conjunto de variáveis de controle e variável de desempenho, o tempo de atravessamento,

e suas relações causais.

A escolha de técnicas é o método monoquantitativo, quando uma técnica de coleta de dados é usada em estudos quantitativos. A proposta desta tese foi usar os dados sobre características processuais (variáveis qualitativas) para propor um modelo sobre o tempo de tramitação processual. As técnicas propostas são: máquina de vetor suporte para classificação, redes neurais e máquina de vetor suporte para regressão e análise de sobrevivência.

A análise de sobrevivência foi escolhida por ser uma técnica estatística habitual para análise de dados de tempo, com objetivo de predição e análise das variáveis influenciadoras: Moore (2016) define análise de sobrevivência como o estudo de tempos de sobrevivência e os fatores que o influenciam. De acordo com Allison (2010), um dos objetivos da análise de sobrevivência é o de estimar modelos casuais ou preditivos nos quais o risco de um evento depende das covariáveis e, segundo Colosimo e Giolo (2006), o conceito da técnica está associado ao tempo, pois a resposta é longitudinal por natureza. Desta forma, análise de sobrevivência pode ser considerada a técnica usual em estudos de tempo.

A pesquisa realizada foi longitudinal retrospectiva, considerando que a variável é o tempo até a ocorrência de um evento de interesse, no caso deste estudo, a baixa do processo. A proposta é fazer um corte no tempo e retroagir para o passado, ou seja, processos baixados em 2017, investigando as variáveis desde a sua autuação no Tribunal Regional Federal da 4^a Região. Segundo Saunders, Lewis e Thornhill (2009), a principal característica da pesquisa longitudinal é a capacidade de estudar a mudança e desenvolvimento. Desta forma, para a análise de sobrevivência, não há censura nos dados.

As propostas de técnicas e procedimentos empregados na pesquisa são redes neurais para regressão, máquina de vetor suporte para classificação e regressão (as três de mineração de dados) e análise de sobrevivência. Esse tema é detalhado na próxima seção.

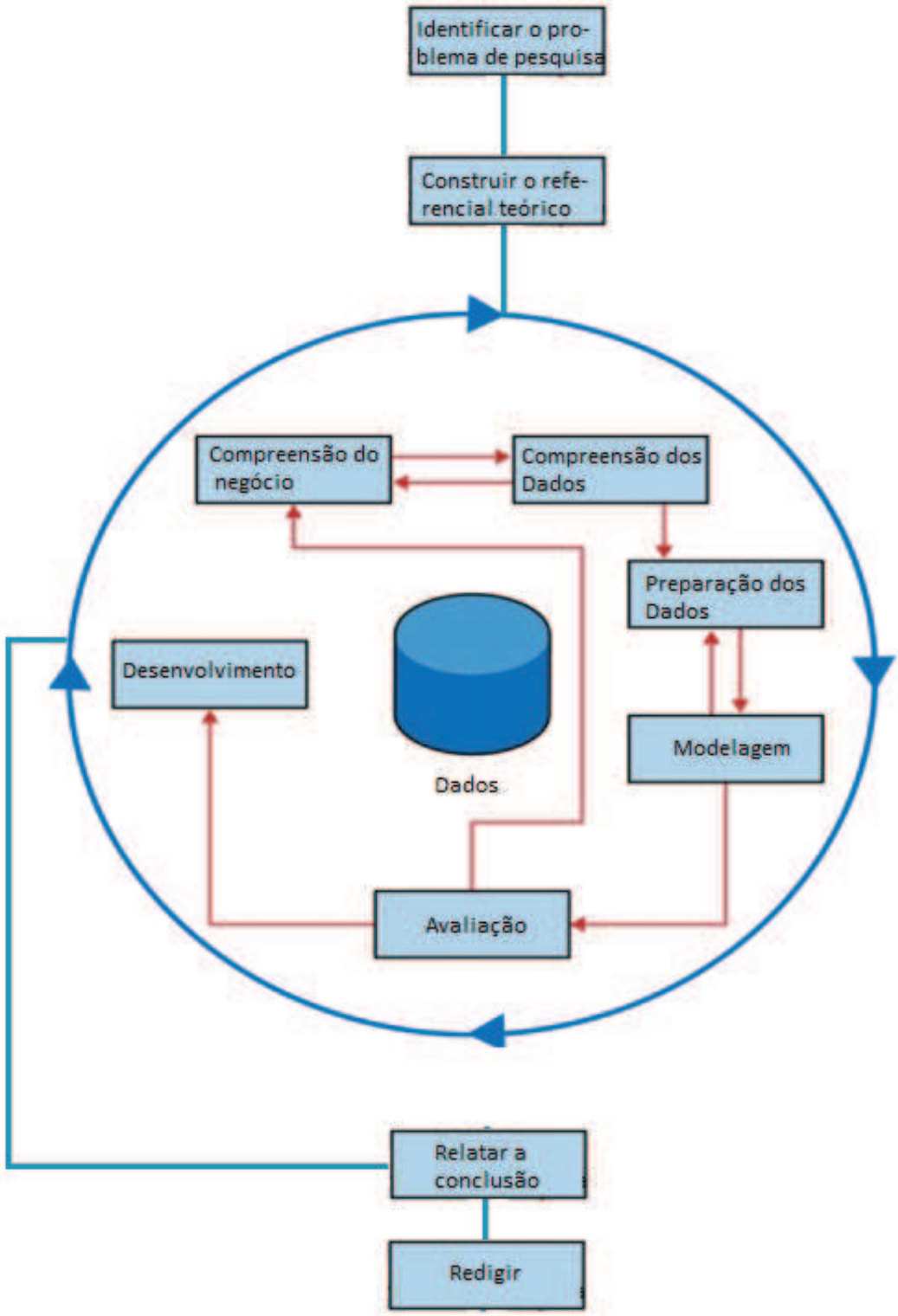
4.2 Método de trabalho

O método de trabalho proposto está na Figura 13. O método é baseado, em parte, no método CRISP, *Cross-industry Standard process for Data Mining*, como está em (NORTH, 2012); (PROVOST; FAWCETT, 2013). O método de trabalho foi escolhido porque entre as as técnicas propostas, redes neurais, máquina de vetor suporte para regressão e classificação são classificados como aprendizado de máquina e mineração de dados. As fases estão descritas a seguir.

4.2.1 Identificar o problema de pesquisa

Após o reconhecimento da importância das instituições públicas, entre elas as organizações do judiciário, para a sociedade e economia, foram buscados, em março de 2016, em

Figura 13 – Método de trabalho



Fonte: A Autora

artigos acadêmicos e notícias, os problemas do judiciário, tanto brasileiro como mundial, as suas relações e propostas de melhoria.

Foi realizada uma busca no Google acadêmico sobre problemas do judiciário, a qual resultou em 359 trabalhos, independente do ano. No portal de periódicos da Capes, a busca foi feita pelas palavras-chaves *Court problems*, resultando em 17 artigos referentes a crimes.

Foi procurado, no saite do CNJ, sobre problemas do judiciário, resultando em 46 notícias. A partir dos três principais problemas reconhecidos do judiciário, conforme CNJ (2014), a busca seguiu concentrada no problema da morosidade, especificamente, sobre o tempo de atravessamento no judiciário brasileiro, principalmente, em trabalhos quantitativos sobre este tema.

Simultaneamente, foram levantados os programas de gestão pública, a forma de medir esta morosidade e a importância dada para o tema nos programas de gestão pública. Com base neste levantamento, constatou-se a relevância da morosidade processual e sua forma de medir.

Até março de 2016, havia 456 trabalhos no google acadêmico com as palavras chaves "tempo processual", "Brasil" e "judiciário". Estes trabalhos estavam relacionados, principalmente, ao direito e à gestão. Contudo, até então, havia 8 pesquisas quantitativas realizadas no judiciário brasileiro, conforme descritos no Capítulo 1. No Google, buscou-se pelas palavras judiciário, *Data Mining* e tempo, retornando 88.000 resultados.

No portal de periódicos da Capes, foram buscadas as palavras chaves *time*, *lawsuit* e *Data Mining*, esta última representando a forma quantitativa, porém, não apresentando resultados nos últimos 10 anos. Da mesma forma, foram pesquisadas pelas palavras chaves *time*, *lawsuit* e *Survival Analysis*, também não mostrando resultados.

Na busca por trabalhos acadêmicos empregando métodos quantitativos à duração processual, foi pesquisado, na base de dados da EBSCO, em março de 2017, os termos *length of judicial proceedings*, *court delay*, *disposition time*, *filings court time* e *time to court case resolution* acrescido de *data mining*, *regression analysis*, *survival analysis*, *support vector machine*, *random forest* e *naive Bayes*, sem filtros, como ano de publicação. Os resultados estão na Tabela 3.

Tabela 3 – Trabalhos realizados sobre tempo processual

Palavras	Support de Vector Machine	Naive Bayes	Random Forests	Data Mining	Regression Analysis	Survival Analysis
Length of judicial proceedings	0	0	0	2	4	0
Court delay	0	0	0	2	19	3
Time to Disposition time and e court	0	0	0	1	51	2
Filings court time	0	0	0	0	0	0
Time to Court Case Resolution	0	0	0	0	0	0

Fonte: A Autora

Conforme reportado no Capítulo 1, os modelos de predição pesquisados na literatura empregaram, principalmente, a análise de regressão e redes neurais, como os trabalhos de Dalton e Singer (2014) e Pavanelli, Pavanelli e Costa (2013), respectivamente. Outras técnicas podem ser propostas para análise diferente das aplicadas. As formas possíveis identificadas são análise de sobrevivência e máquina de vetor suporte. Além disso, os trabalhos encontrados ou diferenciavam os processos criminais dos demais ou tratavam por assuntos específicos.

Por outro lado, foi disponibilizado acesso ao *datawarehouse* do TRF4, às variáveis disponíveis, possibiliatando a realização de um estudo sobre tempo de atravessamento processual.

Da busca em trabalhos acadêmicos constatou-se que, em gestão judiciária, uma área da gestão pública brasileira, há poucos trabalhos sobre modelos de previsão de tempo de atravessamento processual. Ainda não há trabalhos sobre o tempo mensurado em escala categórica no judiciário. Na justiça federal, especificamente, ainda não há publicações sobre modelos de previsão de tempo de atravessamento.

Para contribuir com os estudos nesta área, esta tese pretende, conforme relatado no Capítulo 1: fazer uma revisão dos trabalhos publicados a respeito do tempo de atravessamento processual; propor modelos de previsão deste tempo utilizando outras técnicas quantitativas de análise; e analisar o tempo por meio de faixas.

A partir da exploração inicial do tema, da identificação do problema de pesquisa, da descrição dos trabalhos realizados, a próxima etapa foi a construção do referencial teórico.

4.2.2 Construir o referencial teórico

O passo seguinte foi a construção do referencial teórico abordando a estrutura do judiciário brasileiro e da Justiça Federal da 4^a Região, bem como mineração de dados e análise de sobrevivência para servir de apoio à resolução do problema de pesquisa.

A descrição do judiciário e da justiça tem o propósito de contextualizar a instituição e os problemas contemporâneos. Também foi realizada uma busca por trabalhos abordando a modelagem ou predição do tempo processual, principalmente, no judiciário brasileiro. Os resultados estão apresentados no Capítulo 2. A construção do referencial auxiliou, principalmente, no cumprimento do objetivo específico de discutir os modelos quanto a tipo de estudo, a variável resposta utilizada e covariáveis usadas.

4.2.3 Compreensão do negócio ou compreensão organizacional

De acordo com North (2012), o passo de compreensão do negócio ou compreensão organizacional é fundamental para um resultado de mineração bem sucedido, e consiste em definir as questões a serem respondidas. Para Provost e Fawcett (2013), como o processo de mineração de dados é um ciclo, isto é, se a formulação inicial não for completa, várias iterações podem ser necessárias até o surgimento de uma solução aceitável. Segundo os autores, a equipe pode pensar no uso de cenários.

A compreensão do negócio está relacionada no Capítulo 2, quando foi descrita a organização da justiça brasileira, a Justiça Federal, a competência da Justiça Federal e outros conceitos relacionados. Neste momento é necessário responder questões específicas para atingir os objetivos específicos e geral da pesquisa. Estas questões são:

- Todas as variáveis devem ser incluídas no modelo?
- Quais variáveis influenciam o tempo de atravessamento? O Quadro 2 mostrou os principais trabalhos tratando de tempo de atravessamento no Poder Judiciário Brasileiro, de acordo com as variáveis usadas (etapa de compreensão organizacional) e a técnica usada (etapa de modelagem)? Esta questão foi complementada com entrevistas a especialistas.
- Poder haver recodificações ou transformações nas variáveis?

4.2.4 Compreensão dos dados

O passo seguinte, a compreensão dos dados, diz respeito a uma atividade preparatória consistindo em juntar, identificar e compreender seus ativos de dados, porque, segundo North (2012), a presença de dados imprecisos ou incompletos é indesejável em uma atividade de mineração de dados. Provost e Fawcett (2013) complementam a ideia afirmando que conhecer os dados é importante para compreender sua força e limitações, considerando que raramente os dados ajustam-se aos problemas. Os autores ainda observam que os custos dos dados variam, pois há os disponíveis virtualmente enquanto outros são dispendiosos para obter ou inexistentes, sendo a estimação dos custos e os benefícios das fontes de dados os pontos mais críticos nesta etapa.

O *datawarehouse* do TRF4 está sendo implantado e a proposta é de integrar os módulos judicial, recursos humanos, administrativo e financeiro. Atualmente, o módulo judicial está

bastante avançado, sendo que os demais módulos estão em fase de implantação. A autuação dos processos no TRF4 segue a Resolução 441 do Conselho da Justiça Federal (CJF). A resolução resolve que os registros referentes aos processos, no momento da distribuição, devem observar a classificação e a codificação da Tabela de Classes, de Assuntos e de Entidades da Justiça Federal.

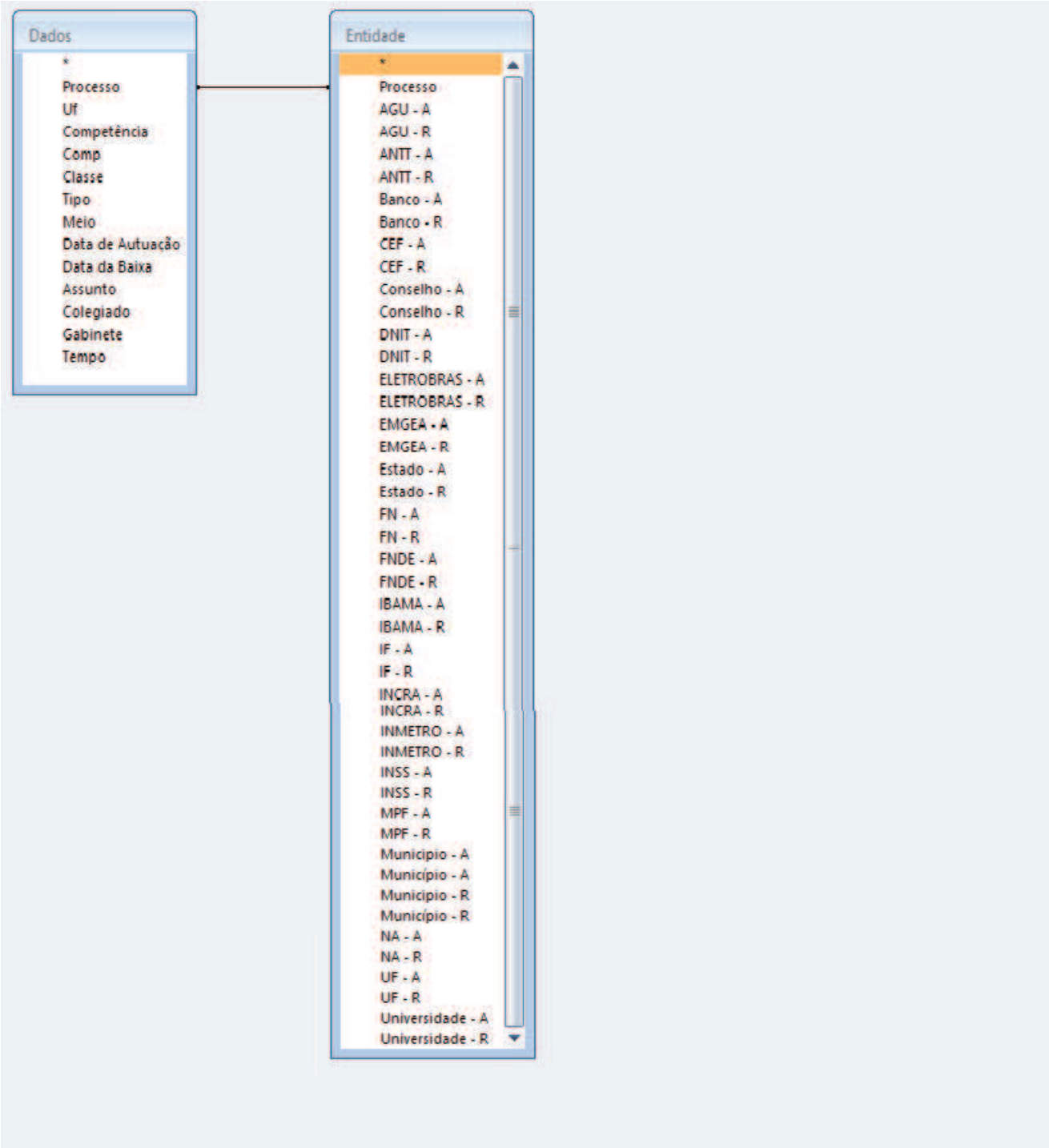
Como a proposta é realizar a análise por meio de um conjunto grande de dados já disponível, não foram propostas outras variáveis, ausentes nos bancos de dados. Estas variáveis só poderiam ser inseridas no estudo mediante coleta, ou manual ou por *web scraping* (extração de dados mediante a utilização de rastreadores *bot* em que informações específicas são coletadas da web), por amostragem e/ou por tema de pesquisa. De qualquer forma, esta coleta seria onerosa em relação ao tempo e recursos empregados. A seguir a descrição das variáveis disponíveis no *datawarehouse* do TRF4:

- Ano: Ano do registro;
- Mês: Mês do registro;
- Data da estatística: Data em que o relatório é gerado;
- Data de autuação: Data de autuação do processo;
- Assunto do processo (níveis): A categorização por assunto é feita pela Tabela Única de Assuntos da Justiça Federal. É uma tabela hierárquica, em até cinco níveis. O primeiro nível apresenta o ramo do direito: Direito Administrativo, Direito Civil, Direito Tributário, Direito Previdenciário, Penal, do Consumidor, do Trabalho, Processual do Civil e do Trabalho, Processual Penal, Marítimo, Marítimo e Internacional. Os níveis seguintes são pormenorizações do assunto principal. Como exemplo de um assunto em cinco níveis pode ser citado o Direito Tributário (nível 1), Crédito Tributário (nível 2), Extinção do crédito tributário (nível 3), interrupção (nível 4) e despacho de citação (nível 5). Há assuntos que podem ser bem descritos já no segundo nível, como por exemplo, Direito Tributário (nível 1), Dívida Ativa (nível 2);
- Entidade: como a Justiça Federal trata de conflitos entre cidadãos e a Administração Púbica Federal, necessariamente, de um lado do conflito, aparecem as empresas públicas como o INSS, autarquias e fundações publicas federais, tais como o IBAMA, os conselhos de fiscalização profissional e a União, etc;
- Pólo: Pólo (Ativo ou Passivo) da entidade. Um único processo pode ter mais de uma entidade e assim, o Pólo pode ser ativo e passivo simultaneamente;
- Evento: Evento registrado no processo, de acordo com a TUMP – Tabela Única de Movimentação Processual;
- Sistema: Sistema fonte (SIAPRO, E-procV2, respectivamente para processos físicos ou eletrônicos);

- Meio: Meio do processo (físico, eletrônico). Esta variável é igual a sistema e uma das duas pode ser suprimida da análise;
- Advogado/Procurador (Nome): Nome do Advogado/Procurador;
- OAB: Número na OAB do Advogado/Procurador;
- Advogado/Procurador (A/P): Categoria (Advogado, Procurador);
- Gabinete: Gabinete do relator;
- Magistrado: Magistrado relator do processo. Esta variável, como a anterior, pode mudar ao longo do tempo;
- Colegiado: Turma ou Seção do processo: são 8 turmas e 4 Seções. Esta variável tem relação com o assunto (nível 1) do direito e a classe do processo;
- Competência 2^o Grau: Competência do processo (Tributário, Administrativo, etc.). Esta variável tem base no assunto e classe do processo;
- Data da carga (2^o Grau): Data em que é realizada a carga no sistema;
- Estatística (2^o Grau): Relatório para obter as variáveis quantitativas (Distribuição, Tramitação, etc.);
- Classe (2^o Grau): A classificação por classes é feita de acordo com a Tabela Única de Classes da Justiça Federal. As classes identificam o procedimento indicado na petição inicial e tem por objetivo padronizar a terminologia entre os órgãos do Poder Judiciário.
- Número do processo de 2^o grau: Número do processo no registro do 2^o grau;
- Tipo de Justiça: Órgão de origem do processo (Justiça Federal, Estadual, Eleitoral, do TRF, etc.);
- Município: Município de origem do processo. É o município da Comarca ou Vara original do processo;
- UF: UF de origem do processo.

A Figura 14 mostra as variáveis do *datawarehouse* do TRF4. As variáveis quantitativas são a quantidade de processos e o tempo em dias entre determinadas fases. A variável quantidade de processos, geralmente, está relacionada à variável qualitativa estatística de 2^o Grau, definindo as diversas variáveis quantitativas: entre outras, distribuídos, julgados, baixa, tramitação, etc. As variáveis de quantidade de processos (quando relacionadas à variável 'Estatística') podem ser: situacionais, ou seja, representam a quantidade de processos em uma data de referência (por exemplo, a quantidade de processos em tramitação em 31 de dezembro de 2014); ou cumulativas, representando a quantidade de processos acumulada em um período de tempo (por exemplo, como a distribuição de processos). Há dados disponíveis de processos desde 2006.

Figura 14 – Variáveis disponíveis no BI do TRF4



Fonte: A Autora

Entretanto, observa-se que existem variáveis de pouca utilidade para este trabalho, ou complementares a outras: número OAB do Advogado, nome do advogado, categoria Advogado, data da carga, não parecem ser relevantes para esta pesquisa. Com o identificador do advogado, a quantidade de advogados por processo poderia ser obtida e então, usada como variável quantitativa. Porém, esta variável não está disponível para processos antigos, o que inviabiliza a sua utilização.

As variáveis referentes aos tempos processuais são mensuradas em dias com o tempo entre dois marcos referenciais. Atualmente, estas variáveis estão disponibilizadas pela quantidade de processos e tempo médio. Medidas de posição e variabilidade podem ser calculadas listando os processos e os seus tempos. Pode haver o cruzamento entre os tempos e as variáveis qualitativas descritas anteriormente, à exceção de entidade e pólo. Os tempos médios disponíveis são os seguintes:

- Distribuição até Julgamento;
- Baixa em Diligência até Retorno da Diligência Cumprida;
- Distribuição até Análise de Recursos Excepcionais;
- Distribuição até Remessa STJ/STF;
- Análise de Recursos Excepcionais até Remessa STJ/STF;
- Remessa STJ/STF até Baixa;
- Remessa STJ/STF até Retorno STJ/STF;
- Distribuição até Baixa;
- Julgamento até Intimação/Publicação do Acórdão.

As variáveis podem ser alteradas durante o curso do processo, ou por erro de codificação, ou por mudança, no caso das classes ou por inclusão, como no caso das entidades. Todavia, as alterações de codificação erradas são corrigidas diretamente no sistema. Para atingir os objetivos foi realizada uma pesquisa com especialistas sobre as variáveis processuais e a importância das mesmas no tempo de atravessamento, para conhecer e validar a inclusão ou exclusão de variáveis no modelo. Foram escolhidos, por julgamento, 8 respondentes para atribuir o grau de importância às variáveis do banco de dados.

Foi aplicado um questionário fechado em vez de entrevistas pelos seguintes motivos: i) como as variáveis estão disponíveis no *datawarehouse*, já estão listadas, a agregação de outras variáveis poderia ser inviabilizada pelo tamanho da amostra; ii) verificar a possibilidade de exclusão de alguma variável; iii) a verificação do nível da variável assunto. Foram 8 especialistas, servidores lotados na área judiciária do TRF, reconhecidos pela experiência no trâmite judicial.

Pela impossibilidade de coletar, manualmente ou por *web scraping*, variáveis ausentes no *datawarehouse*, a busca foi reduzida a estas variáveis. De qualquer forma estas variáveis são consideradas atributos de entrada (input).

O instrumento de pesquisa foi um questionário contendo perguntas abertas e fechadas. As perguntas sobre o grau de importância das variáveis foi medida em uma escala intervalar de 7 pontos, apresentando em seus extremos (1) Pouco Importante e (7) Muito Importante. O instrumento de pesquisa foi colocado no Google formulários. O apêndice A apresenta uma cópia. A Tabela 4, a seguir, apresenta o resumo das respostas.

Tabela 4 – Grau de importância atribuído às variáveis

Variável	Mínimo	Mediano	Médio	Máximo	Desvio padrão
Classe	6	6	6,4	7	0,49
Assunto	3	6	5,7	7	1,28
Tipo de justiça	1	4	4,0	7	2,00
Unidade da Federação	1	1	1,9	4	1,36
Meio Processual	4	7	6,3	7	1,03
Gabinete	4	7	6,1	7	1,12
Entidade	2	4	4,1	6	1,12

Fonte: Dados da pesquisa

A variável classe processual foi considerada a mais importante, com grau médio de 6,4, enquanto a variável Unidade da Federação, teve o menor grau de importância (1,9) entre as variáveis pesquisadas. A seguir, os argumentos dados para cada uma das variáveis. Os argumentos referentes à variável classe referem-se à:

- complexidade das classes: “uma Apelação Cível é mais complexa que um Agravo de Instrumento, por exemplo.”
- processos de trabalho: “A classe permite regras de automação que permitem atribuir maior celeridade a tipos que demandam maior urgência” e “A classe do processo agrega circunstâncias comuns às ações". A classe no 2^o grau determina a competência do colegiado, então processos de Seção, Corte e Plenário levariam mais tempo tramitando”; “Quando bem trabalhado, bem detalhado, pode facilitar o reconhecimento de precedentes assemelhados, o que auxilia na elaboração de decisões mais céleres e de melhor qualidade.”
- Quanto ao assunto, há argumentos que ressaltam a importância da classificação correta e da classificação de assuntos repetitivos, como pode ser observado pelas seguintes justificativas: “Com a identificação precisa do assunto é possível facilitar triagens e relocalizações de processo, bem como dinamizar a distribuição interna de trabalho", “Como em geral o assunto não é muito bem cadastrado, acho uma variável pouco confiável para justificar o tempo de tramitação do processo”; “Assuntos que já tem decisões tendem a ser mais

rápidos, em que pese a dificuldade que as partes tem em definir o assunto de forma mais específica."

Quando questionado quanto ao nível da tabela de assuntos, todos os entrevistados mencionaram o último nível. É dada a sugestão de classificação de dificuldade do processo: "Somente pelo grau de dificuldade do assunto classificado".

Em relação à origem do processo, esta variável obteve o segundo menor grau médio de importância, 4,0. Para tanto, as justificativas foram sobre a padronização dos processos oriundos da Justiça Federal: "Os processos oriundos da Justiça Federal possuem uma maior padronização de forma e são feitos no sistema eProc"; Por outro lado, a origem do processo não deveria interferir na celeridade: "não é prática usual discriminar as ações por sua origem, apenas quanto ao tempo decorrido desde a distribuição e quanto ao grau de dificuldade do litígio (em razão de estarmos submetidos a regimes de cumprimento de metas, faz-se necessário administrar processos de rápida resolução com processos que demandem mais tempo de exame)".

A importância da variável meio processual pode ser bem exemplificado pela seguinte citação: "Processos físicos tendem a demorar mais por conta das etapas burocráticas de envio e remessa dos autos"; "Além disso, apenas os processos oriundos da Justiça Estadual (competência delegada) ainda são físicos."

A variável Gabinete, com grau médio de importância atribuído de 6,1, pode ser justificada, em resumo, pelas seguintes citações: "São vários fatores inerentes ao gabinete que interferem no tempo de tramitação (ex.: quantidade de servidores, processo de trabalho, acervo existente, etc)", e "A forma de gerenciamento particularizado de cada gabinete interfere muito na produtividade geral".

Entretanto a variável gabinete não foi incluída nos modelos por dois motivos. Primeiro, é possível um processo iniciar em um Gabinete e ser redistribuído para outro(s) ao longo do fluxo processual. Assim, um processo pode ser ligado a um ou mais gabinetes. Porém, em 2017, devido à criação das Turmas Suplementares, este fato foi frequente a ponto de tornar as conclusões sobre o tempo por gabinete equivocada. O segundo motivo é o fato do Gabinete ser nominado pelo Magistrado titular, possibilitando, desta forma sua identificação. Por estes dois motivos, a variável não foi incluída nos ajustes dos modelos.

Em relação às entidades, o grau médio de importância atribuído, igual a 4.1, é justificado pelos mesmos argumentos da unidade da federação, porém, há o argumento em favor das variáveis de entidade de que "há algumas entidades que possuem um corpo de procuradores e assistentes melhor organizado que outras, o que pode influenciar o tempo do processo."

4.2.5 Preparação dos dados

A preparação de dados, segundo North (2012), envolve as atividades de junção de dois ou mais conjuntos de dados, redução dos dados para selecionar apenas as variáveis de interesse,

limpeza (exclusão dos dados incompletos e dos *outliers*). De acordo com Provost e Fawcett (2013), alguma conversão de dados é necessária antes da fase de modelagem para que os dados produzam melhores resultados. Por exemplo, técnicas de mineração de dados requerem os dados em formato categórico, enquanto outras, em formato numérico, devem ser normalizados ou escalonados. Para a análise de sobrevivência, as variáveis categóricas estavam em formato numérico, para o ajuste dos demais modelos, declaradas como fator. Estas variáveis são também chamadas de covariáveis.

Para a análise, tanto descritiva como para o ajuste dos modelos, é tomado o ano de 2017 porque, além de ser o último, representa melhor as mudanças tecnológicas e administrativas implantadas. A quantidade de processos baixados no TRF4 aumenta ano a ano, como pode ser observado na Figura 3. Nos últimos três anos, a quantidade de processos baixados também aumentou: foram baixados, respectivamente, 109.780, 100.753 e 137.621, processos de competência não criminal no TRF4.

A coleta dos dados foi realizada em janeiro de 2018. Inicialmente, a ideia era analisar os dados de 2016, porém, a decisão tomada foi de usar os dados de 2017, por dois motivos: i) obtenção de dados mais atuais; ii) alteração na organização do TRF. Em junho de 2017, foram instaladas duas Turmas Regionais Suplementares, uma em Curitiba e outra em Florianópolis. As Turmas têm competência para o julgamento de recursos em matéria previdenciária e de assistência social originários dos respectivos estados. O objetivo de criação destas Turmas foi conciliar o interesse dos mais necessitados, em especial nas questões previdenciárias, e o pleito da comunidade jurídica de aproximação física.

As variáveis foram coletadas em dois bancos de dados: i) variáveis gerais; ii) entidades. As variáveis foram coletadas nestes dois bancos porque pode haver várias entidades por processo, assim como o Pólo. Além disto, no *datawarehouse*, não era possível, cruzar a variável entidade com as descritas no banco de dados. A relação entre entidade e demais variáveis (classe, assunto, etc.) foi realizada. A seguir a descrição das variáveis coletadas.

- Processo: identificador;
- Unidade da federação: PR, RS, SC e outros;
- Competência: Administrativo (Seção), Administrativo (Turma), Corte Especial, Presidência, Previdenciário (Seção), Previdenciário (Turma), Tributário (Seção), Tributário (Turma), Vice-Presidência. Esta variável foi recodificada para a variável a seguir;
- Comp: Administrativo, Corte Especial, Presidência, Previdenciário, Tributário, Vice-Presidência;
- Gabinete: Os 24 Gabinetes, conforme a Figura 15, a exceção dos penais;

Figura 15 – Organização do TRF4

Composição e Competência das Turmas, Seções, Corte Especial e Conselho de Administração

Carlos Eduardo Thompson Flores Lenz	**Presidente**	
Maria de Fátima Labarrère	**Vice-Presidente**	
Ricardo Teixeira do Valle Pereira	**Corregedor Regional**	
DESEMBARGADOR(A)/ JUIZ(A) FEDERAL	**TURMA**	**COMPETÊNCIA**
Roger Raupp Rios - Pres.	1ª Turma	Trabalhista e Tributária
Juiz Federal Marcelo De Nardi		
Juiz Federal Alexandre Rossato da Silva Avila		
Rômulo Pizzolatti	2ª Turma	
Luciane Corrêa Münch - Pres.		
Sebastião Ogê Muniz		
Marga Inge Barth Tessler	3ª Turma	Administrativa, Civil e Comercial
Rogério Favreto - Pres.		
Vânia Hack de Almeida		
Luis Alberto d'Azevedo Aurvalle - Pres.	4ª Turma	
Cândido Alfredo Silva Leal Junior		
Vivian Josete Pantaleão Caminha		
Luiz Carlos Canalli - Pres.	5ª Turma	Previdência e Assistência Social
Juiz Federal Altair Antonio Gregorio		
Juíza Federal Gisele Lemke		
João Batista Pinto Silveira - Pres.	6ª Turma	
Juíza Federal Taís Schilling Ferraz		
Juiz Federal Artur César de Souza		
Luiz Fernando Wowk Penteado - Pres.	1ª Turma Regional Suplementar do Paraná	Previdência e Assistência Social
Fernando Quadros da Silva		
Juiz Federal Luiz Antonio Bonat		
Paulo Afonso Brum Vaz - Pres.	1ª Turma Regional Suplementar de Santa Catarina	
Celso Kipper		
Jorge Antonio Maurique		
Márcio Antônio Rocha	7ª Turma	Penal
Claudia Cristina Cristofani		
Salise Monteiro Sanchotene - Pres.		
Victor Luiz dos Santos Laus	8ª Turma	
João Pedro Gebran Neto		
Leandro Paulsen - Pres.		

Fonte: TRF4

- Classe: Para a modelagem foram escolhidas as 9 classes mais frequentes: C10828 (Apelação Cível), C10992 (Apelação / Reexame Necessário), C10822 (Agravo de Instrumento), C10855 (Reexame Necessário Cível), C11011 (Ação Rescisória - Seção), C10943 (Ação Rescisória), C10977 (Embargos Infringentes), C11009 (Mandado de Segurança - Turma), C40012 (Pedido de Efeito Suspensivo a Apelação - Turma). A análise descritiva ainda foi realizada com as duas categorias de conflito de competência;
- Tipo: Para a modelagem, apenas as três justiças de origem mais frequentes: Justiça Federal, Justiça Estadual e TRF;
- Meio: físico ou eletrônico;
- Data de Autuação;
- Data da Baixa: A primeira data de baixa do ano de 2017. Há casos em que há levantamento de baixa, ou seja, o processo é baixado, mas ele volta a tramitar e recebe uma nova baixa. Assim, foram desconsiderados os processos que tiveram baixas anteriores a 2017 e, no caso de baixarem em 2017 e voltarem a tramitar, foi considerada apenas a primeira de 2017;
- Assunto: considerado o 5^o nível da tabela de assuntos, conforme resultado da pesquisa;
- Colegiado: 1^a Seção, 1^a Turma, 2^a Seção, 2^a Turma, 3^a Seção, 3^a Turma, 4^a Turma, 5^a Turma, 6^a Turma, 7^a Turma, 8^a Turma, Corte Especial, Plenário, Presidência, Turma Regional Suplementar de Santa Catarina, Turma Regional Suplementar do Paraná, Vice-Presidência;
- Tempo: Data de baixa – Data de autuação +1

As variáveis categóricas classe, assunto, competência, origem do processo foram transformadas em binárias para o ajuste dos modelos. As variáveis meio e entidade já eram binárias. Em comparação à descrição das variáveis, algumas formas retiradas, pelas razões a seguir:

As variáveis ano e estatística foram usadas como filtros: Ano = 2017 e Estatísticas: baixa com decisão, baixa de ofício ou baixa de conciliação. A variável sistema foi retirada por representar a mesma informação da variável meio. A variável município foi retirada por apresentar muitas categorias, conforme apontado nas entrevistas. A data da carga foi retirada por ser uma variável apenas de controle. A variável magistrado foi retirada para não nominar, o que pode ser considerado antiético.

A transformação dos dados foi realizada depois da coleta, utilizando o software R, versões R 3.1.2, 3.4.0 e 3.5.0. R, segundo R Core Team (2015), é um ambiente e linguagem para cálculos estatísticos e gráficos, semelhante à linguagem S, desenvolvida pela Bell Laboratories, sendo um projeto aderente à licença GNU (software livre). De acordo com Han, Kamber e Pei (2011), em transformação, os dados são transformados ou consolidados na forma apropriada para mineração. As estratégias são:

- Alisamento, para remover o ruído, com as técnicas de regressão, *binning* e agrupamento;
- Construção de variáveis, quando novas são construídas e adicionados para auxiliar no processo de mineração;
- Agregação, quando o resumo ou a agregação de operações são aplicadas aos dados.
- Normalização, quando os dados são escalonados em um intervalo menor, tal como -1 a 1 ou 0 a 1;
- Discretização, onde os valores de um atributo numérico são substituídos por intervalos. Por exemplo, para a variável idade, a discretização transforma a variável em intervalos tais como 0 a 10, 11 a 20, etc. ou valores conceituais como jovem, adulto ou sênior;
- Geração de hierarquia de conceito para dados nominais. Por exemplo, variáveis como ruas podem ser generalizados em um nível maior como cidade.

Para esta tese, as variáveis foram transformadas através da discretização, da normalização, da construção de variáveis e da geração de hierarquia de conceito. A construção de variáveis foi usada para a obtenção das variáveis entidade, demais classes, demais assuntos, demais tipo de origem e demais competências, quando foram unidas várias categorias de variáveis em uma única.

Categorias da variável entidade como Universidade Federal foram obtidas a partir da junção de várias Universidades, assim como a entidade município. Como pode haver várias entidades no mesmo processo, após a junção, elas foram classificadas em sim/não.

A geração da hierarquia de conceitos foi utilizada na variável município, a qual foi obtida através de unidade da federação.

A variável resposta foi normalizada. De acordo com Han, Kamber e Pei (2011), a normalização pode ser por escalonamento decimal, min-max ou z score. A normalização por min-max realiza uma transformação linear de tal forma:

$$v_i^{'} = \frac{v_i - min_A}{max_A - min_A} + (max_{novoA} - min_{novoA}) + min_{novoA} \tag{63}$$

onde $v_i^{'}$ é o novo valor, max_A e min_A são os valores de mínimo e maximo para A, max_{novoA} e min_{novoA} são os novos valores de máximo e mínimo para A.

No procedimento de normalização z-score, os valores da variável obtidos na forma:

$$v_i^{'} = \frac{v_A - \mu(A)}{\sigma(A)} \tag{64}$$

$v_i^{'}$ é o novo valor. Uma variação é o desvio médio absoluto de A:

$$v^{'} = \frac{v_i - \mu(A)}{s_A} \tag{65}$$

Onde $s_A = \frac{|v_i - \mu(A)|}{n}$

O desvio médio absoluto *SA* é mais robusto quando há a presença de *outliers* que o desvio padrão, σ, segundo Han, Kamber e Pei (2011).

De acordo com Han, Kamber e Pei (2011), a discretização de dados e a geração de hierarquia também são formas de redução de dados, pois simplifica os dados originais e torna a mineração mais eficiente. Os padrões resultantes extraídos são geralmente mais fáceis de entender.

Técnicas de redução de dados têm como objetivo reduzir o conjunto de dados, mas manter a integridade dos dados originais. Ou seja, a mineração no conjunto de dados reduzido deve ser mais eficiente, mas produzir os mesmos (ou quase os mesmos) resultados analíticos, de acordo com Han, Kamber e Pei (2011). As estratégias incluem redução de dimensionalidade (redução de variáveis aleatórias sob consideração através de métodos como *wavelets* ou análise de componentes principais), redução de numerosidade (substituição do volume original de dados por formas alternativas e menores de representação), compactação de dados e amostragem.

Em redução da numerosidade, as técnicas podem ser i) paramétricas, quando um modelo é usado para estimar os dados, de modo que normalmente apenas os parâmetros de dados precisam ser armazenados, em vez dos dados reais; ii) não paramétricas. Exemplos de técnicas paramétricas são regressão e modelos log lineares, enquanto os não paramétricos incluem histogramas, agrupamento, amostragem e agregação de cubo de dados.

Segundo Han, Kamber e Pei (2011), uma das vantagens da amostragem para redução de dados é o custo menor em relação ao custo de obter a população. A amostragem é mais comumente usada para estimar a resposta a uma consulta agregada, em redução de dados e considerada uma escolha natural para o refinamento progressivo de um conjunto de dados reduzido. Tal conjunto pode ser ainda mais refinado simplesmente aumentando o tamanho da amostra. A amostragem foi utilizada para o refinamento dos modelos de classificação e regressão.

4.2.6 Modelagem, com o objetivo da predição

O passo seguinte, a modelagem, produziu modelos para predição. De acordo com North (2012), é a aplicação dos algoritmos para procurar, identificar e mostrar algum padrão ou mensagem nos dados, sendo que os tipos básicos são a classificação e a predição. Os objetivos da descoberta de conhecimento são definidos pela intenção do uso do sistema: verificação, para confirmar as hipóteses do usuário; ou descoberta, quando o sistema automaticamente descobre novos padrões. A descoberta pode ainda ser subdivida em predição e descrição, segundo Han, Kamber e Pei (2011), onde as tarefas de mineração descritivas caracterizam as propriedades gerais do banco de dados; quanto às preditiva, elas realizam inferências, a fim de fazer previsões.

A modelagem foi feita utilizando o software R. O R apresenta diversos pacotes para as análises propostas:

- para redes neurais artificiais: *neural net*, Fritsch, Guenther e Guenther (2012), *AMORE* Limas et al. (2010) e *deepnet* (Rong (2014));
- para máquinas de vetor suporte (classificação e regressão): *e1071* (Meyer et al. (2015));
- para a análise de sobrevivência: survival (Hastie (2016)) e *flexsurv* (Jackson e Jackson (2017));

Para a análise, os passos seguidos foram:

1. Análise descritiva com o objetivo de descrever e identificar as variáveis qualitativas, identificando as categorias com maior frequência. Para tanto, foi considerado todo o conjunto de dados e os *missing values*;
2. Retirada dos *missing values*;
3. Seleção de amostra. Para identificar os parâmetros para os modelos de regressão de redes neurais e de regressão e classificação para máquina de vetor suporte, foi extraída uma amostra aleatória do conjunto de dados de treinamento, a fim de tornar mais rápida a escolha dos parâmetros dos modelos, considerando o tempo de treinamento. Foi selecionada uma amostra de tamanho 20.000 com a técnica de amostragem aleatória simples sem reposição. Considerando a população de tamanho 89.418, a amostra foi planejada para estimar uma proporção qualquer da população. Para tanto, o erro seria de 0,8%, com um grau de confiança de 99%;
4. Divisão do conjunto total de dados em treinamento e teste, quando 70% do conjunto de teste foi usado para o treinamento e 30%, para teste. No item anterior, o treinamento foi com os dados da amostra;
5. Aplicação de cada uma das técnicas.

Os modelos ajustados, a exceção do modelo de classificação, foram comparados através da raiz quadrada média do erro de predição, RMSE, para mostrar o quanto as predições estão próximas dos valores observados:

$$RMSE = \sqrt{\frac{(predito - observado)^2}{n^2}} \tag{66}$$

onde n é o número de observações. No caso de redes neurais e máquina de vetor suporte para regressão, para o cálculo do RMSE, os valores preditos foram transformados para a escala do tempo original. O Quadro 5 apresenta as etapas seguidas para obter os modelos utilizando as diversas técnicas.

Quadro 5 – Etapas para a geração dos modelos

Análise de Sobrevivência	Máquina de vetor suporte (classificação)
Obtenção do estimador de Kaplan Meyer Ajuste de curvas de distribuição Verificação do ajuste das curvas Escolha das covariáveis Ajuste do modelo de regressão Verificação do modelo Cálculo do RMSE	Categorização para o tempo Escolha da função kernel Escolha dos parâmetros Obtenção do modelo Aplicação do modelo aos dados de teste Obtenção da matriz de confusão Cálculo de medidas de desempenho Avaliação do modelo
Redes neurais	Máquina de vetos suporte (regressão)
Escolha e transformação da variável resposta Escolha do algoritmo Escolha dos parâmetros Obtenção do modelo Aplicação do modelo aos dados de teste Transformação dos valores preditos para a escala original Cálculo de RMSE (na escala original) Avaliação do modelo	Escolha e transformação da variável resposta Escolha da função kernel Escolha dos parâmetros Obtenção do modelo Aplicação do modelo aos dados de teste Transformação dos valores preditos para a escala original Cálculo de RMSE (na escala original) Avaliação do modelo

Fonte: A Autora

Para a análise de sobrevivência, o estimador de Kaplan Meyer foi obtido para a função de sobrevivência do total e para as covariáveis classe, assunto, competência, entidade, meio, origem e unidade da federação.

Para o ajuste de curvas de distribuição, da análise de sobrevivência, foram propostos os modelos exponencial, Weibull e lognormal. Os pacotes usados para o ajuste foram o *survival*, destinado à análise de sobrevivência e o *MASS* (base). O pacote *flexsurv*, também para análise de sobrevivência, permite, de acordo com Jackson e Jackson (2017), além das distribuições mencionadas, a Gamma generalizada, a distribuição F generalizada, a Gamma, a Log Logística e Gompertz. Porém, devido ao volume de dados, o pacote *survival* respondeu mais rápido e foi utilizado.

A verificação do ajuste das curvas foi realizado através da comparação gráfica com a função de sobrevivência de Kaplan-Meyer, e a escolha das covariáveis para compor o modelo foi por meio do método *stepwise*.

As opções para o ajuste do modelo são: i) de regressão paramétrico de tempo de vida acelerado; ii) modelo paramétrico de tempos proporcionais e iii) modelo semi-paramétrico de Cox. Para a verificação de adequação do modelo, foram feitos gráficos dos resíduos.

O RMSE para comparação com os demais modelos (de redes neurais e máquina de vetor suporte) foi calculado com base no conjunto de teste.

Foram ajustados dois modelos de regressão, um para redes neurais e outro por meio de máquina de vetor suporte. Para o modelo de redes neurais, os algoritmos testados foram: retropropagação, retropropagação resiliente com e sem retrocesso de peso.

A escolha dos parâmetros do modelo de redes neurais foram o número de neurônios, o valor da taxa de aprendizado, o número de camadas ocultas na rede e o número de iterações como critério de parada.

Tanto para o modelo de máquina de vetor suporte para regressão como o de redes neurais, foram obtidos valores preditos a partir da aplicação do modelo obtido com dados de treinamento aos dados de teste. Uma vez que a variável resposta foi transformada, estes valores preditos estavam em escala diferente da original. Assim, estes valores preditos foram transformados para a obtenção de valores na escala em dias para o calculo do RMSE.

Como pacotes para análise estão o *neuralnet* (Fritsch, Guenther e Guenther (2016)), e o *deepnet* (Rong (2014)). O pacote *neuralnet* tem por foco algoritmos de aprendizagem supervisionados: de acordo com Fritsch, Guenther e Guenther (2012), *neuralnet* foi preparado para treinar perceptrons multi-camada no contexto da análise de regressão, ou seja, redes neurais usadas como extensão de modelos lineares generalizados. O pacote *deepnet*, de acordo com Rong (2014), implementa arquiteturas de *deep learning* e algoritmos de rede neural como retropropagação e RBM (*Restricted Boltzmann Machine*).

Inicialmente, a ideia era utilizar o pacote *neuralnet*, o qual tem por foco algoritmos de aprendizagem supervisionado. Embora fornecendo um bom material de análise, mostrou-se lento para retornar resultados, e foi substituído pelo pacote *deepnet*.

Em redes neurais não há regras claras para encontrar o melhor número de camadas ocultas, bem como os demais parâmetros. Desta forma, o design é um processo de tentativa e erro dependente de valores iniciais tais como os pesos.

Para os modelos de máquina de vetor suporte, tanto para classificação como para regressão, a escolha da função kernel foi entre linear, radial, polinomial e sigmoide. Segundo Meyer et al. (2015), para a função polinomial, é possível escolher o grau, enquanto para as funções radial, polinomial e sigmoide pode-se escolher o γ. Para a função radial, ainda é possível escolher o custo. Para o modelo de regressão de SVM, ainda houve a escolha do tipo de regressão (υ-regressão ou ε).

Com a utilização de máquina de vetor suporte, tanto para regressão como para classificação, foi utilizado o pacote e1071, que utiliza a bibilioteca LIBSVM, proposta por Chang e Lin (2011), a qual realiza a análise em duas etapas: primeiro, treina um conjunto de dados para obter um modelo e segundo, usa o modelo obtido para fazer previsões utilizando um conjunto de dados de teste.

Conforme Meyer e Wien (2001), a biblioteca do R apresenta as seguintes possibilidades com máquinas de vetor suporte:

- υ-classification: permite maior controle sobre o número de vetores de suporte, especificando um parâmetro adicional que se aproxima da fração de vetores de suporte.
- Classificação de uma classe: permite a detecção de *outliers*/novidades;
- Classificação multi-classe;
- ε -regressão;
- υ-regressão.

Meyer e Wien (2001) fornecem as seguintes dicas tanto para classificação como para regressão:

- Verificar um intervalo de combinações de parâmetros.
- Para classificação, pode ser usado o kernel RBF (padrão), devido ao seu bom desempenho geral e poucos números de parâmetros (apenas dois: C e γ). Os autores sugerem tentar valores pequenos e grandes para C como de 1 a 1000 primeiro, e em seguida, decidir quais são os melhores valores de Cs para os dados por validação cruzada. Só então tentar vários γ para os melhores valores de Cs;
- No entanto, os autores recomendam usar a função tune.svm no pacote e1071. A função tune, de acordo com Meyer et al. (2015), ajusta os hiper parâmetros de métodos estatísticos usando uma pesquisa de grade sobre os intervalos de parâmetros fornecidos;
- Os tempos de treinamento aumentam rapidamente com um grande volume de dados;
- Escalonar os dados.

Para obter o modelo utilizando máquina de vetor suporte para classificação, foi necessário, inicialmente, categorizar os dados, isto é, transformá-los, da escala em dias para faixas de tempo. Neste modelo, é possível escolher os pesos para as classes.

Diferente dos outros três modelos, para a classificação, as medidas calculadas para verificar o desempenho do modelo foram: acurácia, sensibilidade, especificidade, precisão e F1, também a partir dos dados de teste. Isto é, após a aplicação do modelo obtido com os dados de treinamento aos dados de teste, valores preditos foram obtidos. Estes valores foram comparados, por meio de tabelas, aos valores originais de tempo do conjunto de teste e então, as medidas foram calculadas.

Para encontrar os melhores parâmetros do modelo, para os de redes neurais e de máquina de vetor suporte, várias combinações de parâmetros foram testadas e comparadas através das medidas de desempenho. A descrição está no Capítulo 5. Esta combinação foi obtida com os dados da amostra de tamanho igual 20.000.

Após a geração dos modelos, foi aplicado o algoritmo *apriori* para obter regras de associação. O objetivo foi identificar as características dos processos mais rápidos e mais

demorados, para cumprir o objetivo específico de analisar o tempo de atravessamento processual em relação às variáveis que o influenciam. Para obter estas regras foi utilizado o pacote *arules* do R (HAHSLER et al., 2018).

4.2.7 Avaliação

Após a construção do modelo e antes da apresentação dos resultados, é realizada a etapa de avaliação, com a revisão dos passos anteriores. Esta etapa consiste na verificação se a informação é relevante, sendo possível retornar aos passos anteriores, caso os objetivos não tenham sido alcançados de forma satisfatória.

Segundo Provost e Fawcett (2013), consiste em analisar os resultados rigorosamente e garantir que eles sejam válidos e confiáveis antes de avançar no ciclo. Para Zumel, Mount e Porzak (2014), avaliar um modelo é quantificar sua performance. De acordo com North (2012), a avaliação pode ser dada através de uma série de técnicas de validação cruzada, teste para falsos positivos e, também, avaliação incluindo algum aspecto humano. Enfim, a avaliação inclui tanto avaliação quantitativa como qualitativa. Segundo Zumel, Mount e Porzak (2014), quando se constrói um modelo para fazer predições, além da construção, é necessário testar se o modelo faz predições corretas com novos dados. Neste sentido, há um conjunto de dados chamado treinamento, ou o conjunto que alimenta a construção do modelo, e outro, chamado de teste, que é o conjunto de dados que, com os parâmetros estimados pelo conjunto de treinamento, verifica se o modelo de previsão é acurado.

Zumel, Mount e Porzak (2014) aconselham dividir os dados em treinamento e teste antes da modelagem, não comparando com o teste até a avaliação final. Caso seja necessário calibrar, pode ser feito com uma nova divisão dos dados de treinamento. Alternativamente, segundo Zumel, Mount e Porzak (2014), outras técnicas de validação podem ser usadas, como *fold cross validation* ou validação baseada em *bootstrapping*.

Entre as técnicas de análise podem ser citadas *bootstrapping* e *cross fold validation*. Em *bootstrapping*, segundo Zaki e Jr (2006), a ideia é gerar k amostras de um conjunto D de dados utilizando amostragem com substituição. Dado uma estatística de teste θ, a técnica permite obter um intervalo de confiança para θ com um nível de confiança α desejado. Por exemplo, *bootstrapping* pode ser usado para testar se um padrão X é frequente na população com suporte $sup(X)$.

Fold Cross-Validation, de acordo com Zaki e Jr (2006), divide o conjunto de dados D em k partes iguais (*folds*), $D1, D2, ..., Dk$. onde cada conjunto de treinamento é o conjunto de dados menos D_i. O modelo ajustado ao conjunto de dados menos D_i é então aplicado ao conjunto D_i. Para cada D_i, obtém-se uma estatística de θ_i. No final, então, pode ser obtida uma medida de interesse como a média.

Especificamente sobre a avaliação de modelos de classificação, podem ser destacadas

as medidas de acurácia, precisão, recall, F1, sensibilidade e especificidade. Para calcular estas medidas é necessário obter uma matriz de confusão. A Tabela 5 apresenta esta matriz.

Tabela 5 – Tabela de valores preditos e observados

Preditos	Observado	
	Positivo	Negativo
Positivo	Verdadeiros positivos	Falsos positivos
Negativo	Falsos negativos	Verdadeiros negativos

Fonte: A autora

De acordo com Zumel, Mount e Porzak (2014), os itens são classificados nas categorias positivo e negativo e então são obtidos os seguintes totais:

- Total de itens positivos classificados corretamente ou verdadeiro positivos;
- Total de itens negativos classificados como positivos ou falso positivos;
- Total de itens negativos classificados corretamente ou verdadeiro negativos;
- Total de itens positivos classificados como negativos ou falso negativos.

Com estes valores, é possível calcular as medidas citadas. Acurácia, segundo Zumel, Mount e Porzak (2014), é a quantidade de itens classificados corretamente em relação ao total de itens. De acordo com a autora, esta medida não é indicada quando o evento é raro na população, ou tem distribuição ou custos desequilibrados.

As medidas de precisão e recall são calculadas conjuntamente. Precisão, de acordo com Zumel, Mount e Porzak (2014), pode ser obtida pela divisão da quantidade de itens positivos classificados corretamente dividido pelo número de itens classificados como positivos. Recall, conforme Zumel, Mount e Porzak (2014), é a fração de itens positivos detectados corretamente pelo classificador em relação a todos os itens positivos, ou seja, precisão é uma medida de confirmação e recall, de utilidade. Com estas duas medidas, pode-se calcular a medida F1, através da seguinte combinação:

$$F1 = \frac{2 \times precisao \times recall}{precisao + recall} \tag{67}$$

Sensibilidade e especificidade, segundo Zumel, Mount e Porzak (2014), são medidas de efeito: a sensibilidade é igual à medida recall, enquanto a especificidade é a fração de itens negativos detectados corretamente pelo classificador em relação a todos os itens negativos. Quando a tabela é de ordem $m \times n$, as categorias são unidas com o objetivo de obter uma tabela 2×2. Para os modelos de classificação, a avaliação foi feita através de *cross fold validation*, também.

4.2.8 Desenvolvimento

O passo de desenvolvimento, segundo Provost e Fawcett (2013), envolve atividades de automatização do modelo, reuniões com os consumidores, integração com os sistemas de informação, alimentando um novo aprendizado do uso do modelo para melhorar sua precisão e desempenho, monitoramento e medição das consequências do uso deste modelo.

Desta forma, de acordo com Han e Kamber (2006), a mineração de dados é apenas um passo de todo o processo. Embora essencial e considerando que o termo mineração de dados está se tornando mais popular que o termo *Knowledge Discovery from Data*, ele insere-se em um framework analítico maior. Segundo Provost e Fawcett (2013), o ciclo CRISP-DM gira em torno de exploração; isto é, sobre as abordagens e estratégias em vez de projetos de software, onde os resultados de um passo podem mudar a compreensão fundamental do problema.

Após a validação dos modelos, pode-se proceder à disponibilização. O conhecimento gerado é, então, apresentado à Organização, compondo a Gestão da Informação, e as fases anteriores são documentadas.

Ou seja, é a apresentação formal aos gestores da empresa, apresentando tanto o modelo quanto as demais análises, explicando as variáveis, as suas relações e, principalmente, como o modelo pode auxiliar na tomada de decisão. Também serão relatados os problemas enfrentados na execução, ressaltando os aspectos positivos de todo o processo.

Desta forma, após a construção dos modelos, um deles pode ser disponibilizado para consulta das partes processuais, que, a partir dos valores das covariáveis, pode-se predizer um valor de tempo de atravessamento.

A realização destas etapas permitiu a construção dos quatro modelos a serem descritos no Capítulo 5, cumprindo os objetivos geral e específicos propostos no Capítulo 1.

5 IMPLEMENTAÇÁO E ANÁLISE

Este capítulo traz, em um primeiro momento, a análise descritiva do tempo de atravessamento e das covariáveis. Após, a descrição dos modelos ajustados e a avaliação de cada um. Por último, as comparações e discussões pertinentes.

5.1 Apresentação dos dados e análise descritiva

A análise refere-se ao tempo de atravessamento dos processos baixados em 2017. No TRF4, foram remetidos para baixa 144.727 ações naquele ano. Destes processos, 7.106 eram criminais e foram retirados do conjunto de dados. Das 137.621 ações de competências não criminais restantes, foram retirados os processos que já haviam tido pelo menos uma baixa antes de 2017, restando 135.258 casos, dos quais realizou-se a análise.

A seguir, são analisadas as covariáveis disponíveis por meio das distribuições de frequência. Após, apresenta-se a análise do tempo de atravessamento, seu histograma e as medidas descritivas.

5.1.1 Variáveis categóricas

A análise concentra-se nas variáveis categóricas (covariáveis) de meio processual, unidade da federação de origem, competência, justiça de origem do processo, classe, assunto, gabinete e entidades. As figuras 16 a 22 mostram os gráficos da distribuição de frequência destas variáveis. Como muitas vezes estas covariáveis estão em sigla ou código, a legenda é apresentada no apêndice B.

Contudo, há valores perdidos provenientes de falta de registro, não mostrados nos gráficos devido à sua baixa frequência. A quantidade de observações válidas está na Tabela 6.

Tabela 6 – Quantidade de processos analisados segundo as variáveis

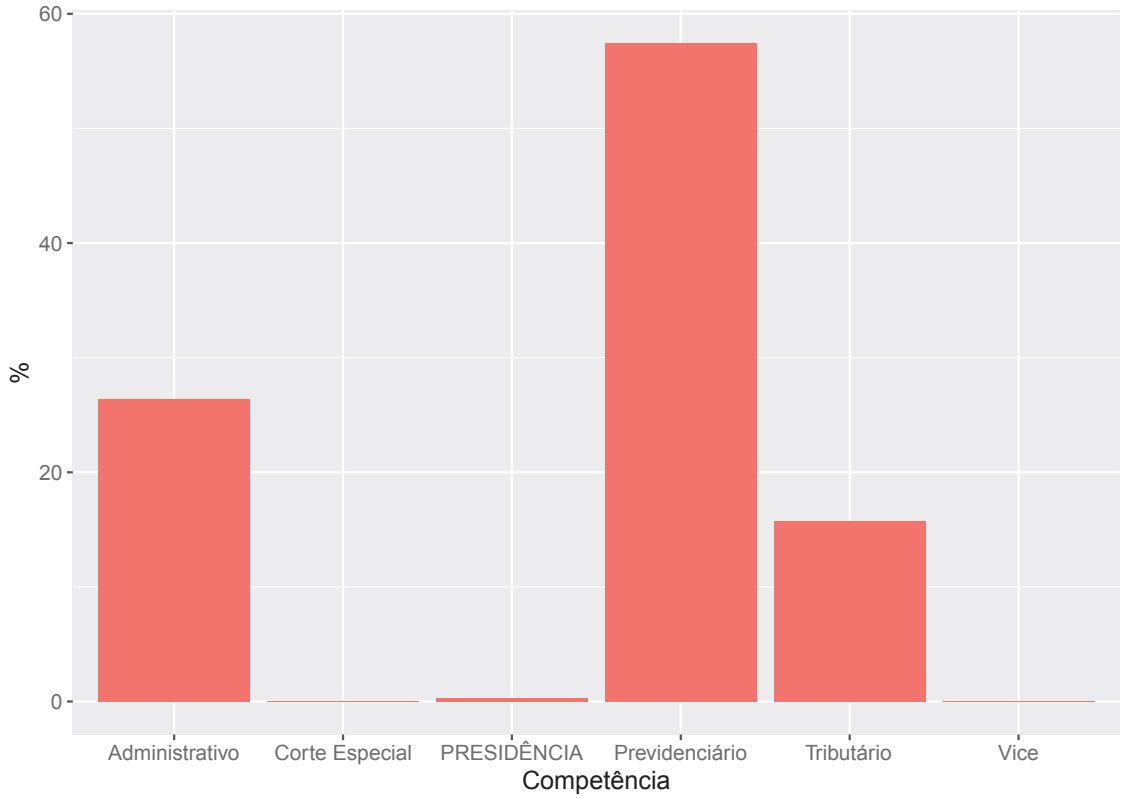
Variável	Observações válidas	Observações perdidas	Total
Meio processual	135258	0	135258
Competência	135258	0	135258
Unidade da Federação de origem	134735	523	135258
Justiça de origem	135189	69	135258
Classe	135212	46	135258
Assunto	135211	47	135258
Entidade	128103	7155	135258

Fonte: Dados da pesquisa

Os gráficos mostram as categorias mais frequentes das variáveis com o objetivo de selecioná-las para o ajuste dos modelos. Como exemplo, pode ser citada a variável assunto, que possui 914 categorias, 115 delas com uma única observação.

A Figura 16 mostra os processos de acordo com as competências. A competência previdenciária é a mais frequente, com 57,4%, seguida da competência administrativa, com 26,4% dos processos e tributária, com frequência de 15,9%. As categorias administrativo, tributário e previdenciário representam, juntas, 99,6% do conjunto de dados.

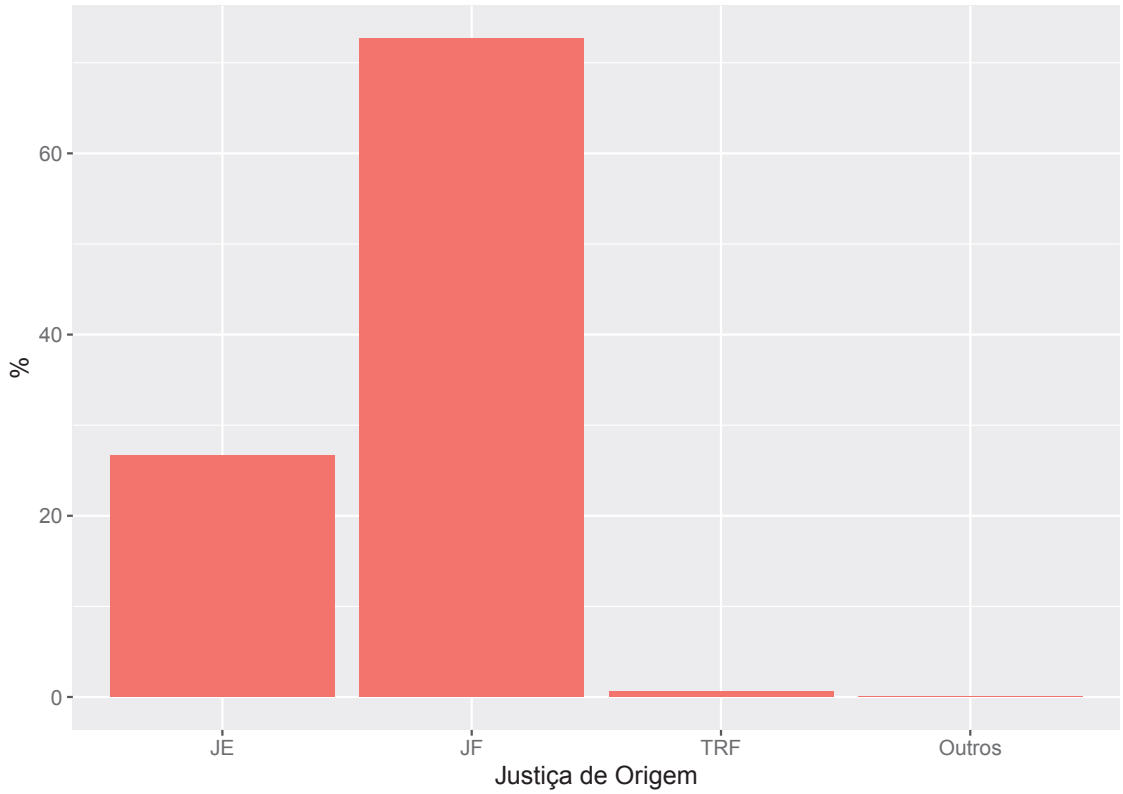
Figura 16 – Distribuição de frequência - Competência processual



Fonte: Dados fornecidos pelo TRF4

Estas três categorias permaneceram na análise. As competências Presidência, Corte Especial e Vice-Presidência foram unidas em uma categoria denominada Demais. Em relação à origem do processo, no gráfico da Figura 17, as categorias Justiça Estadual, Justiça Federal e Tribunal Regional Federal somadas representam 99,9% das observações. A Justiça Federal é a mais frequente, representando 72,6% das observações. Em seguida, há a categoria Justiça Estadual, com percentual de 26,7% das ações. A terceira categoria mais frequente é TRF4, com o equivalente a 0,6% dos processos. As demais categorias desta variável (Justiça do Trabalho, Outras, Tribunal Federal de Recursos, Tribunal de Justiça e Tribunal do Trabalho) foram agrupadas em Demais.

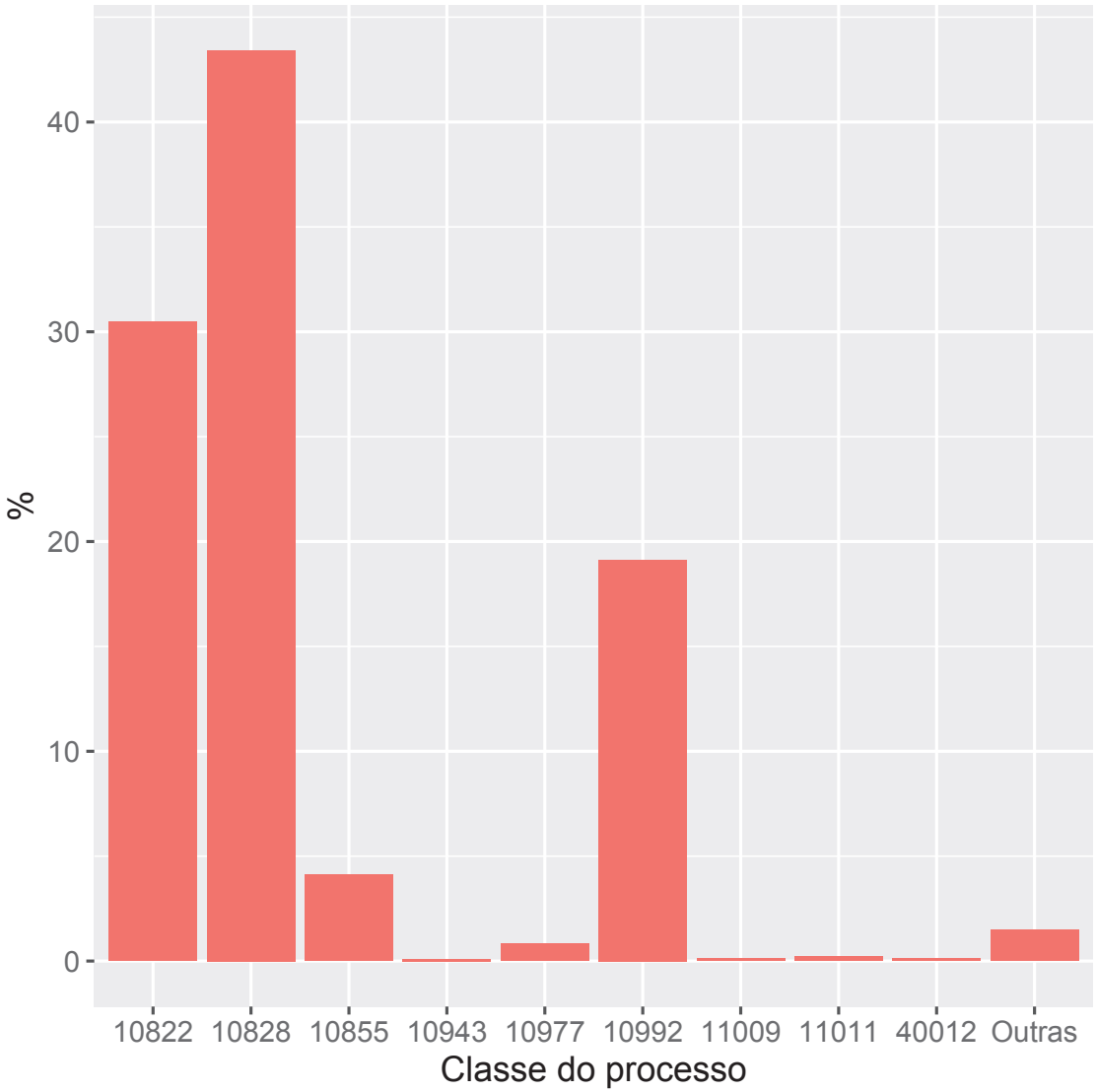
Figura 17 – Distribuição de frequência - Justiça de Origem



Fonte: Dados fornecidos pelo TRF4

Quanto às classes, constantes no gráfico da Figura 18, as 9 classes somadas representam 98,5% dos casos. A classe apelação civil (C10828) é a mais frequente, representando 43,4%dos processos baixados em 2017. A segunda classe com maior baixa de processos é a classe agravo de instrumento (C10822), com 30,5% de processos. A Classe apelação/reexame necessário (C10992) representou 19,1% dos processos.

Figura 18 – Distribuição de frequência - Classe

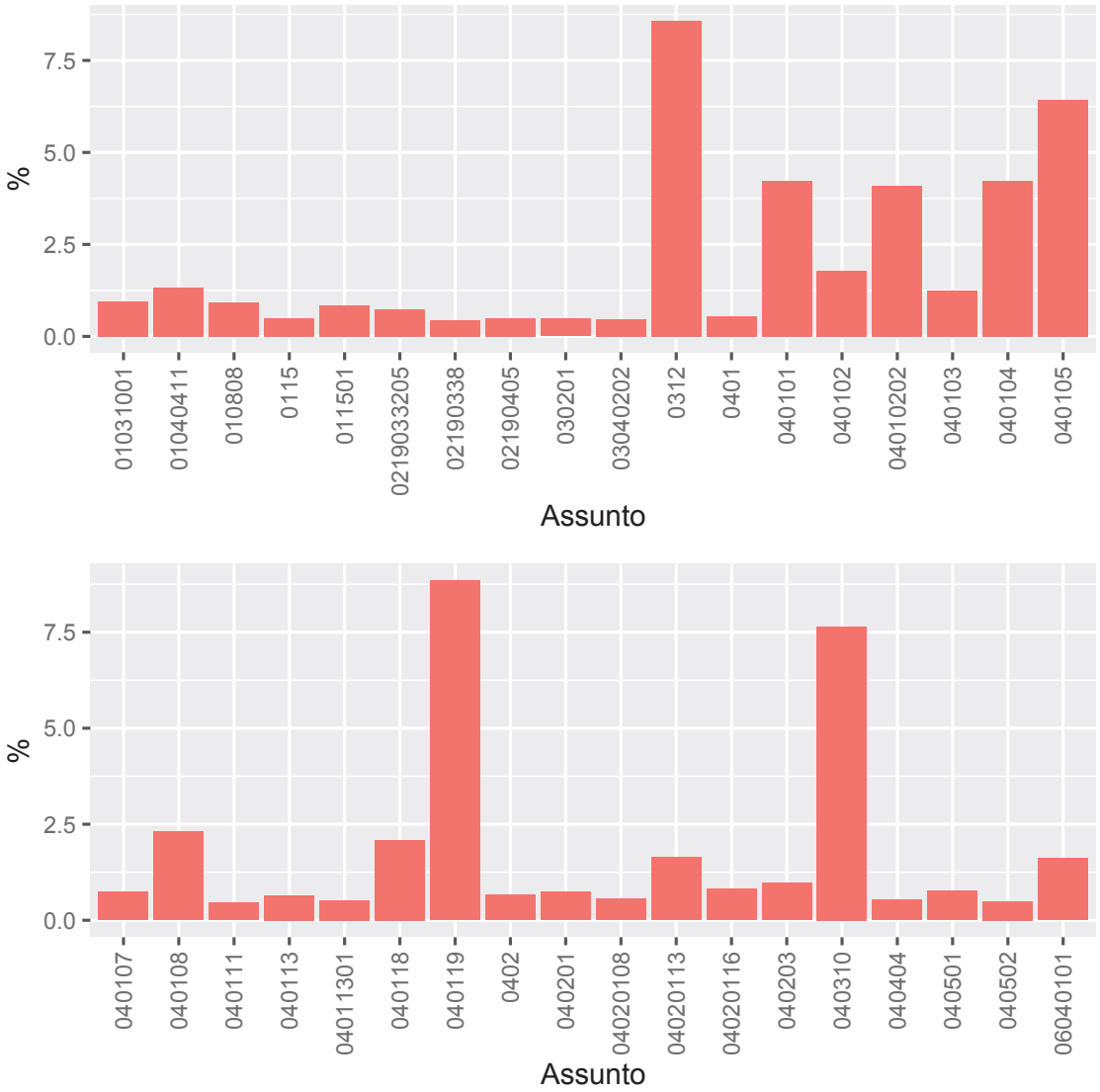


Fonte: Dados fornecidos pelo TRF4

Para o prosseguimento da análise, as classes seguintes foram mantidas: C10828 (Apelação Cível), C10992 (Apelação / Reexame Necessário), C10822 (Agravo de Instrumento), C10855 (Reexame Necessário Cível), C11011 (Ação Rescisória - Seção), C10943 (Ação Rescisória), C10977 (Embargos Infringentes), C11009 (Mandado de Segurança - Turma), C40012 (Pedido de Efeito Suspensivo á Apelação - Turma). As 54 restantes, foram agrupadas na categoria Demais.

Os assuntos estão na Figura 19, classificados conforme o quinto nível da tabela única de assuntos - TUA. Os 36 com maior frequência, entre as 914 categorias de assunto, representaram 70,30%.

Figura 19 – Distribuição de frequência - Assunto do processo



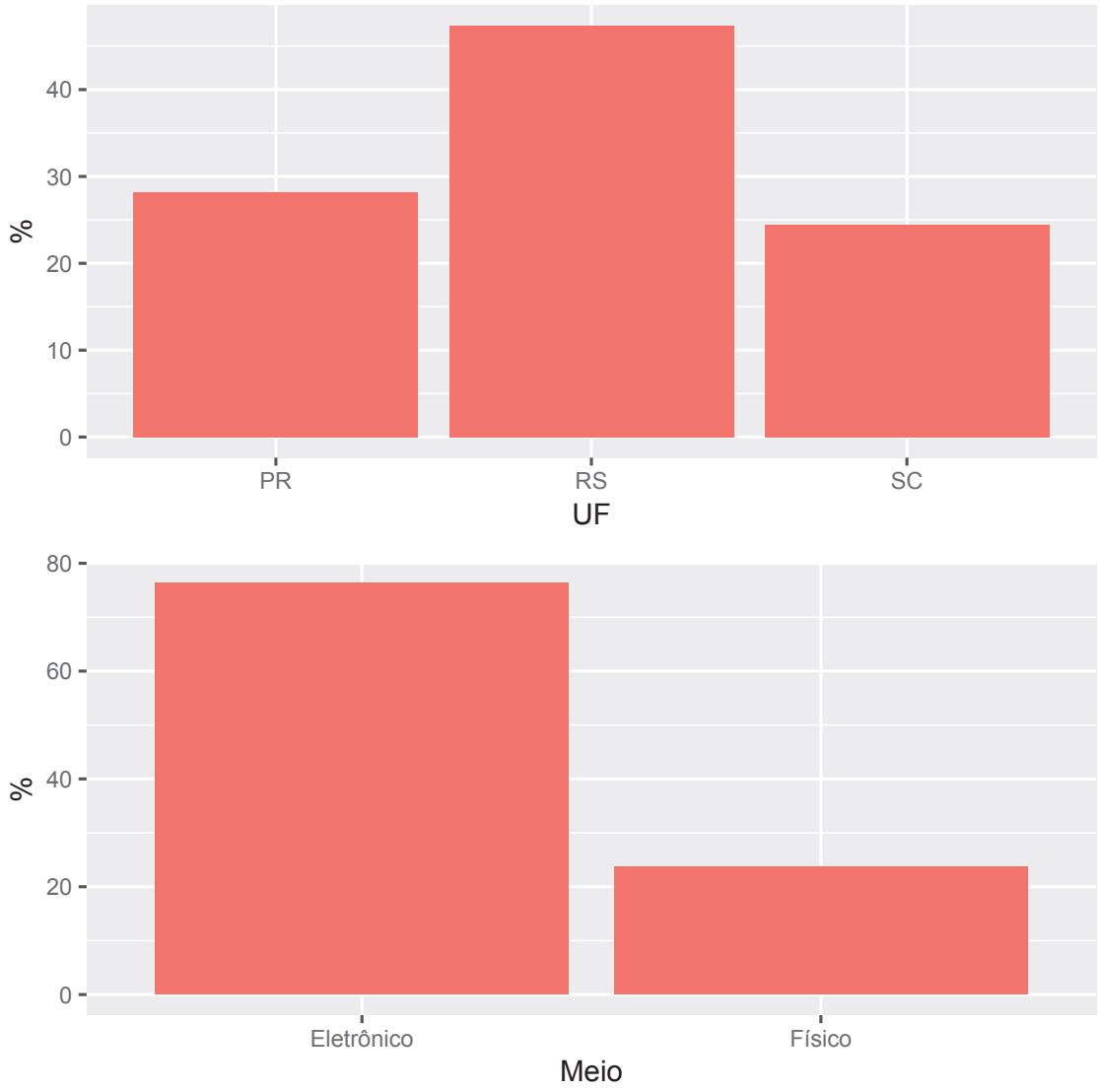
Fonte: Dados fornecidos pelo TRF4

Estes assuntos são: N040119 (Aposentadoria por Tempo de Contribuição (Art. 55/6)), N0312 (Dívida Ativa), N040310 (Renúncia ao benefício), N040105 (Auxílio-Doença Previdenciário), N040101 (Aposentadoria por Invalidez (Art. 42/7)), N040104 (Aposentadoria Especial (Art. 57/8)), N04010202 (Aposentadoria por Idade - Rural (art. 48/51)), N040118 (Aposentadoria por Tempo de Serviço (Art. 52/4)), N040102 (Aposentadoria por Idade (Art. 48/51)), N04020113 (IRSM de Fevereiro de 1994(39,67%)), N06040101 (Expurgos Inflacionários / Planos Econômicos), N01040411 (Fornecimento de Medicamentos), N040103 (Aposentadoria por Tempo de Serviço (Art. 52/6) e/ou Tempo de Contribuição), N040203 (Reajustes e Revisões Específicos), N01031001 (Multas e demais Sanções), N010808 (Seguro-desemprego), N011501 (Multas e demais Sanções), N04020116 (Alteração do coeficiente de cálculo do benefício), N040501 (Averbação/Cômputo/Conversão de tempo de serviço especial), N040201 (RMI - Renda Mensal Inicial), N040107 (Salário-Maternidade (Art. 71/73)), N0219033205 (Seguro),

N0402 (RMI - Renda Mensal Inicial, Reajustes e Revisões Específicas), N040113 (Benefício Assistencial (Art. 203,V CF/88)), N04020108 (Limitação do salário-de-benefício e da renda mensal inicial), N0401 (Benefícios em Espécie), N040404 (Concessão), N04011301 (Pessoas com deficiência), N02190405 (Cédula de crédito rural), N0115 (Dívida Ativa não- tributária), N040502 (Averbação/Cômputo de tempo de serviço de segurado especial (regime de economia familiar)), N030201 (IRPF/Imposto de Renda de Pessoa Física), N03040202 (Cofins), N040111 (Auxílio-Acidente (Art. 86)), N02190338 (Contratos Bancários). Os demais foram agrupados na categoria Demais.

O meio processual e a unidade da federação de origem estão no gráfico da Figura 20. Os processos eletrônicos representaram 76,3% dos processos baixados em 2017.

Figura 20 – Distribuição de frequência - Unidade da federação e meio processual



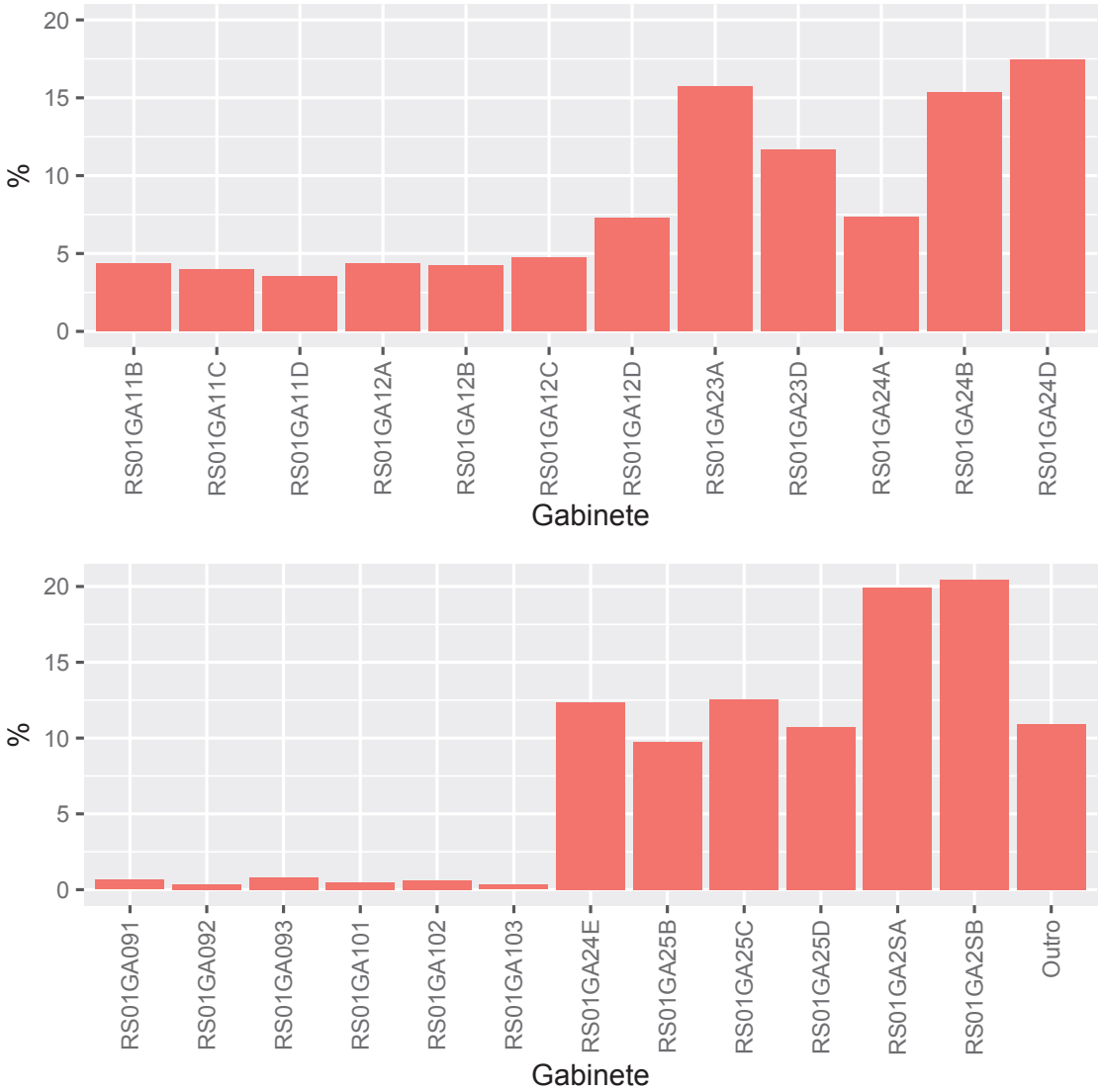
Fonte: Dados fornecidos pelo TRF4

Como atualmente, apenas são distribuídos processos físicos de competência previdenciária oriundos da Justiça Estadual, os físicos referem-se ou a processos antigos das competências tributário e administrativo ou a processos previdenciários.

Quanto à unidade da federação, os processos oriundos do Rio Grande do Sul representaram 47,2% dos processos, os do Paraná, 28,1% e os de Santa Catarina, 24,3%. Nesta variável há 0,3% de processos com perda de registro.

A quantidade de processos baixados por gabinete está na Figura 21. Estes são os 24 gabinetes ativos no final do ano e, embora apareçam codificados, eles são nominados de acordo com o Magistrado titular.

Figura 21 – Distribuição de frequência - Gabinete

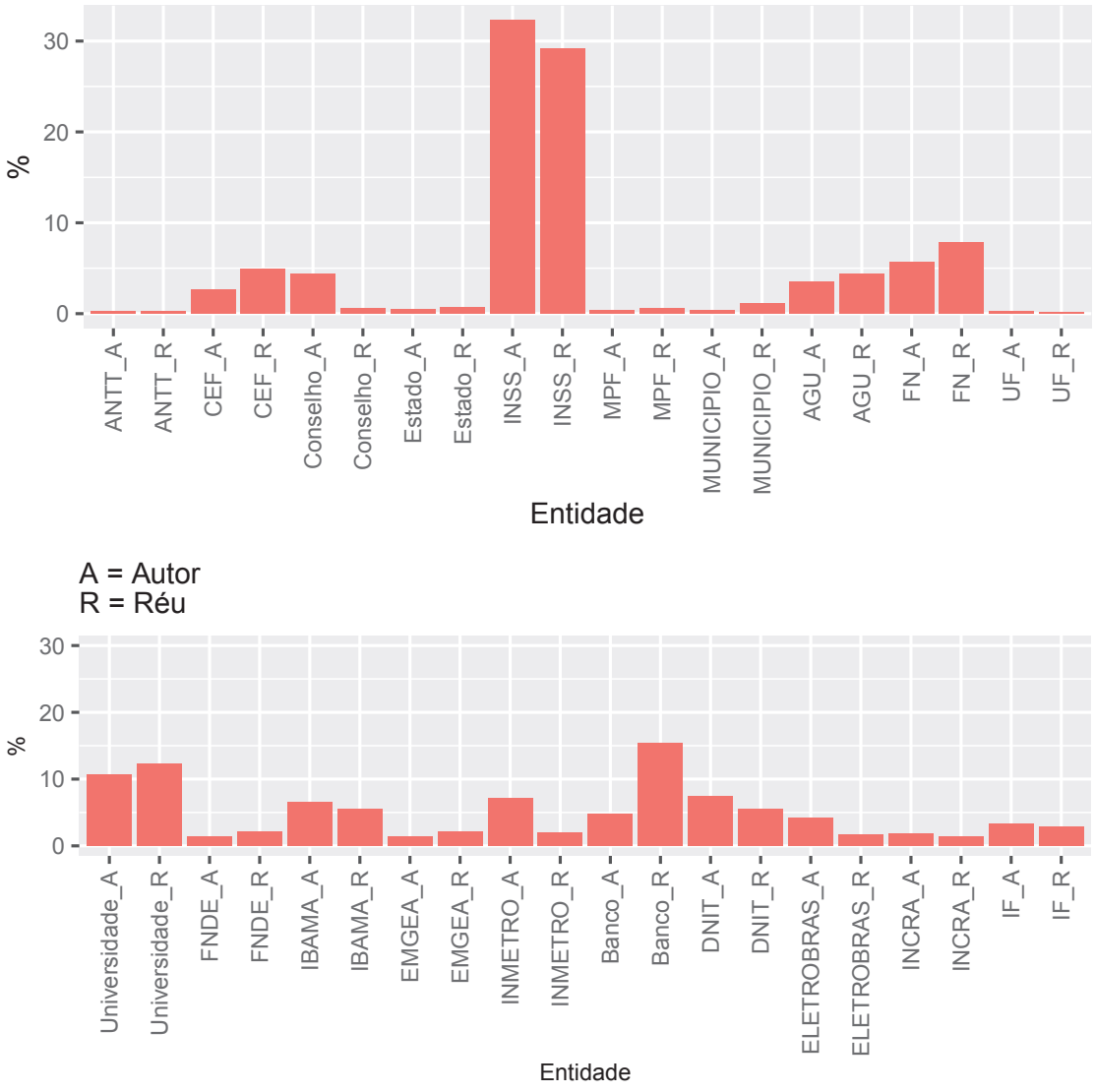


Fonte: Dados fornecidos pelo TRF4

Contudo, 6 destes gabinetes foram implantados ao longo de 2017, mediante a instalação das Turmas Regionais Suplementares, explicando, desta forma, o baixo número de processos baixados por estes gabinetes. Conforme reportado no Capítulo 4, a variável foi retirada do restante da análise para evitar a nominação do magistrado.

Assim como os assuntos, apenas as entidades com maior frequência estão no gráfico da Figura [1]22. As 20 entidades com maior frequência somam 97,0% das observações com informação sobre entidades, apresentadas conforme o pólo, em ré e autora.

Figura 22 – Distribuição de frequência - Entidades



Fonte: Dados fornecidos pelo TRF4

[1]

As entidades com maior frequência são: ANTT (réu e autor), CEF (réu e autor), Conselhos Regionais e Federais (réu e autor), Estado (réu e autor), INSS (réu e autor), MPF (réu e autor), Município (réu e autor), AGU (autor e réu), Fazenda Nacional (autor e réu), União Federal (autor e réu), Universidades Federais (autor e réu), FNDE (autor e réu), IBAMA (autor e réu), EMGEA (autor e réu), INMETRO (autor e réu), Banco (autor e réu), DNIT (autor e réu), ELETROBRAS (autor e réu), INCRA (autor e réu), Instituto Federal (autor e réu). Os processos com entidades diferentes foram agrupados na categorias Demais entidades (ré e autora).

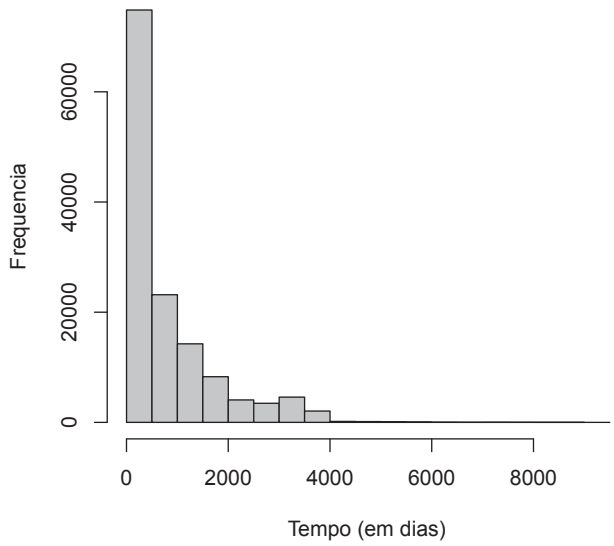
As variáveis classe, meio processual, unidade da federação, assunto, competência e entidades contavam com 102 categorias, incluindo os estratos Demais para classe, assunto, origem e entidades. Assim, houve a tentativa de reduzir a dimensão original dos dados, mas mantendo o máximo da variância original mediante uma Análise de Componentes Principais (ACP), descrita no Apêndice E.

Caso fossem utilizados componentes, em vez das variáveis originais, seriam pelo menos 52, explicando 67% da variância. Assim, pelo alto número de variáveis e baixa explicação da variância original, a opção foi a de manter o número original de variáveis.

5.1.2 Análise do Tempo de Atravessamento

O tempo dos processos cíveis baixados em 2017, mensurado em dias, apresentou média de 790,7 e desvio padrão de 920,13. A mediana, de 400 dias, indica uma distribuição assimétrica à direita, como pode ser observado na Figura 23.

Figura 23 – Histograma do tempo de atravessamento



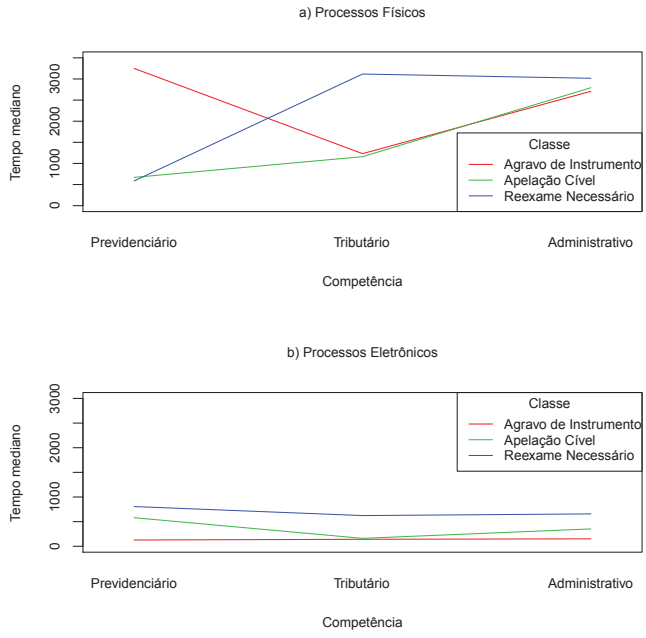
Fonte: Dados da pesquisa

Em relação aos quartis, o valor do 1^o quartil é 149 dias e o 3^o, 1.103. O valor do 95^o percentil é de 3.021 dias, o que indica que os 5% dos processos mais morosos apresentaram tempo superior a este valor.

Quanto às covariáveis, o Apêndice C apresenta a comparação das categorias das variáveis quanto ao tempo através de boxplots. Devido à elevada quantidade de covariáveis, não foram mostradas as interações, isto é, quando a categoria de uma variável depende da categoria de outra.

A título de ilustração, a figura seguinte mostra o efeito das variáveis competência, meio e classe no tempo. O gráfico a) da Figura 24 mostra a interação entre classe e competência para processos físicos. A classe apelação cível, em verde, apresentou tempos medianos menores para processos previdenciários. Isto se deve ao fato que apenas esta competência ainda recebe processos físicos, fazendo com que os processos administrativos e tributários físicos sejam os antigos, necessariamente. O tempo mediano dos agravos de instrumento de processos eletrônicos é praticamente o mesmo para as três competências, respectivamente, de 151, 127 e 142 dias para processos administrativos, previdenciários e tributários.

Figura 24 – Interação entre classe e competência



Fonte: Dados da pesquisa

As seções seguintes apresentam os modelos ajustados: primeiro, a análise de sobrevivência e após, redes neurais, máquina de vetor suporte para regressão e classificação, respectivamente. A estimação das funções de sobrevivência foram realizadas com a totalidade dos dados. Para o ajuste dos modelos, de sobrevivência e os demais, o conjunto de dados com o total de observações válidas foi dividido em dois, 89.418 (70%) para o treinamento e 38.322 (30%) para o teste,

conforme as seções seguintes.

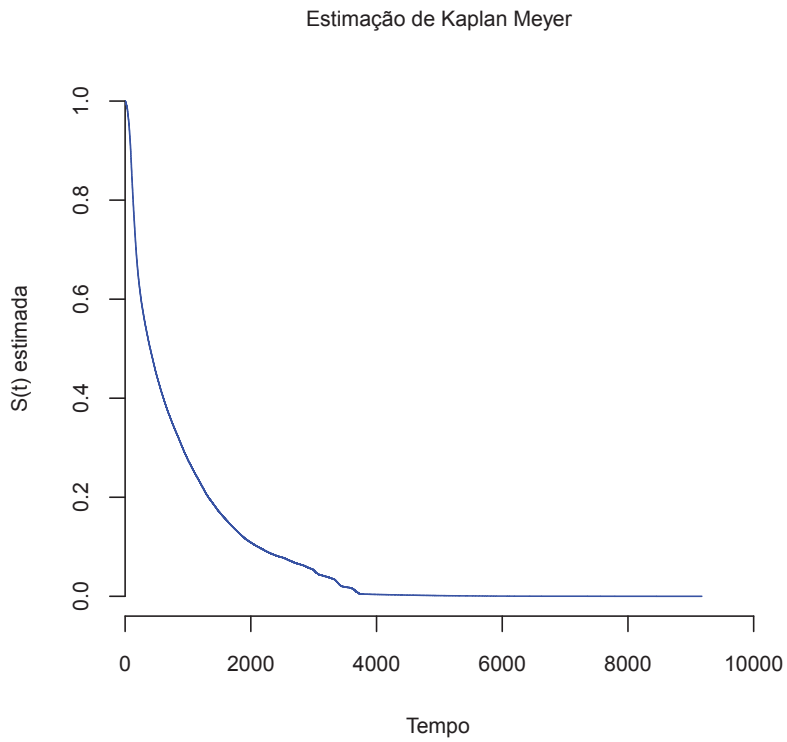
5.2 Análise de sobrevivência

Inicialmente, apresenta-se a estimação das funções de sobrevivência de Kaplan Meyer e a verificação do ajuste de uma distribuição aos dados. Após, a escolha das covariáveis, o ajuste do modelo, a interpretação dos coeficientes e a verificação do ajuste por meio da análise dos resíduos.

Nesta tese a análise de sobrevivência, considerada como técnica habitual para análise de dados de tempo, cumpre dois objetivos: i) analisar as covariáveis e sua influência no tempo de atravessamento e ii) comparar os resultados obtidos com as regressões por redes neurais e máquina de vetor suporte. Em comparação a redes neurais, a análise de sobrevivência analisa sob o ponto de vista das covariáveis e sua importância no tempo do processo.

Assim, como início da análise, mostra-se o estimador de Kaplan Meyer para a função de sobrevivência do conjunto de dados, conforme a Figura 25. O gráfico apresenta a probabilidade de um processo não ter sido baixado no tempo t. Como não há censura nos dados, o estimador é uma função de sobrevivência empírica.

Figura 25 – Função de Sobrevivência

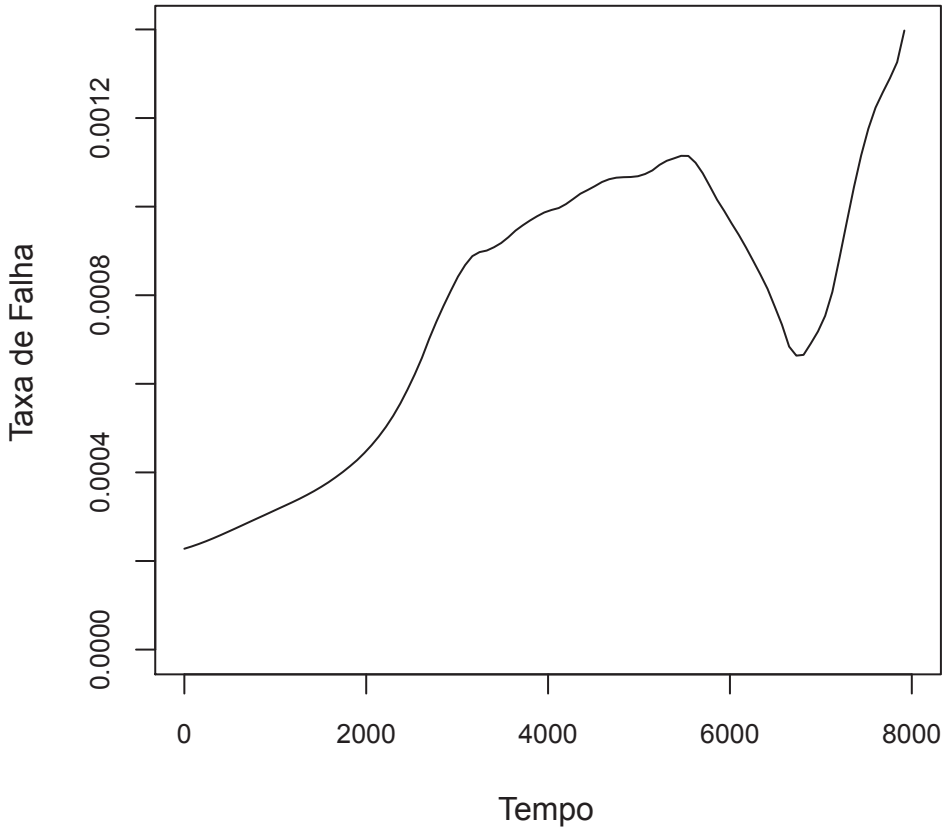


Fonte: Dados da pesquisa

Para o tempo do total de processos, a probabilidade decresce rapidamente até 3500 dias, quando a probabilidade de não ter sido baixado passa a ser 0,0189. A partir deste valor, a probabilidade é próxima a zero. O Apêndice D contém os gráficos do estimador de Kaplan Meyer para as variáveis competência, tipo, origem do processo, classe, unidade da federação e assunto.

A seguir é apresentado o gráfico da função de risco. A taxa de risco estimada apresenta o seguinte comportamento: crescente inicialmente até aproximadamente 5.500 dias. A partir deste valor, decresce até aproximadamente 6.600 dias e volta a crescer, conforme a Figura 26.

Figura 26 – Função de risco estimada

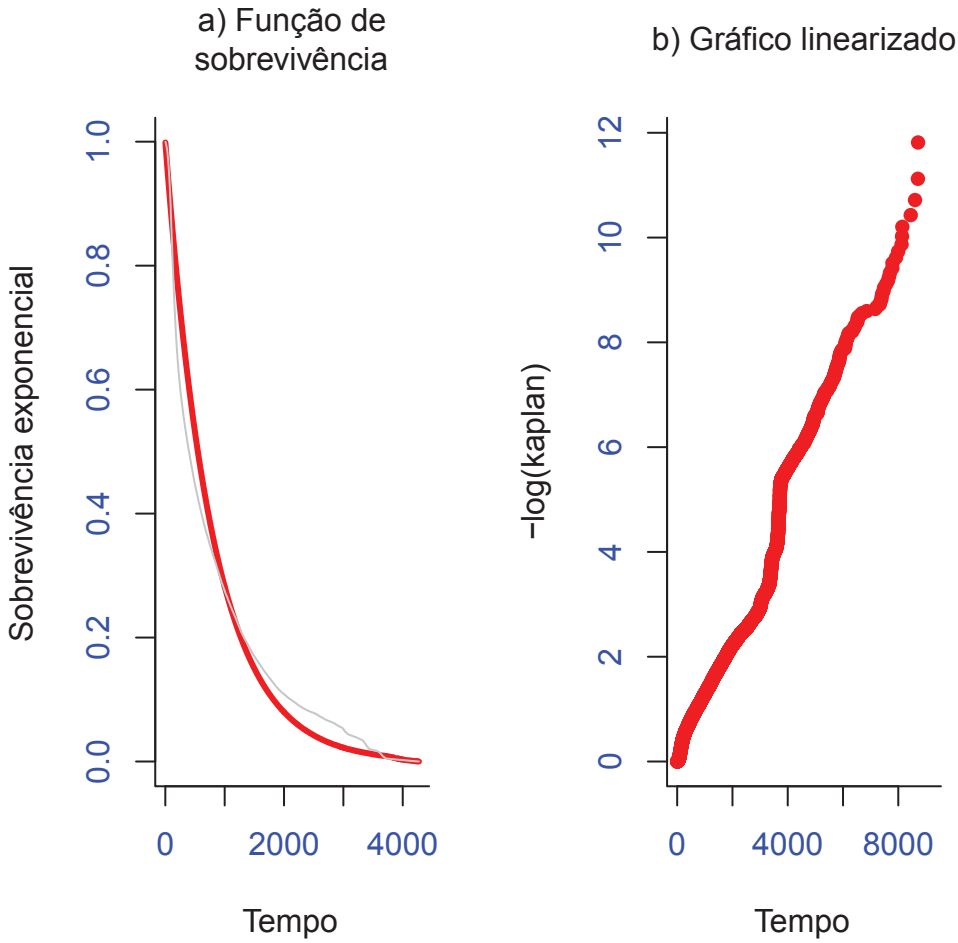


Fonte:Dados da pesquisa

Uma vez estimada as funções de sobrevivência de Kaplan Meyer e a de risco, foram ajustadas curvas de distribuição de modelos exponencial, Weibull e lognormal aos dados. Foram feitas duas comparações para verificar visualmente a adequacidade das distribuições: i) as funções de sobrevivência de acordo com as distribuições ajustadas versus a sobrevivência de Kaplan Meyer e ii) as funções linearizadas.

O gráfico a) da Figura 27 apresenta a função de sobrevivência de acordo com a distribuição exponencial ajustada e o gráfico b), a função linearizada. No gráfico a), a partir do tempo de aproximadamente 2.000 dias, a distribuição afasta-se da exponencial teórica, na linha vermelha.

Figura 27 – Comparação com a distribuição exponencial

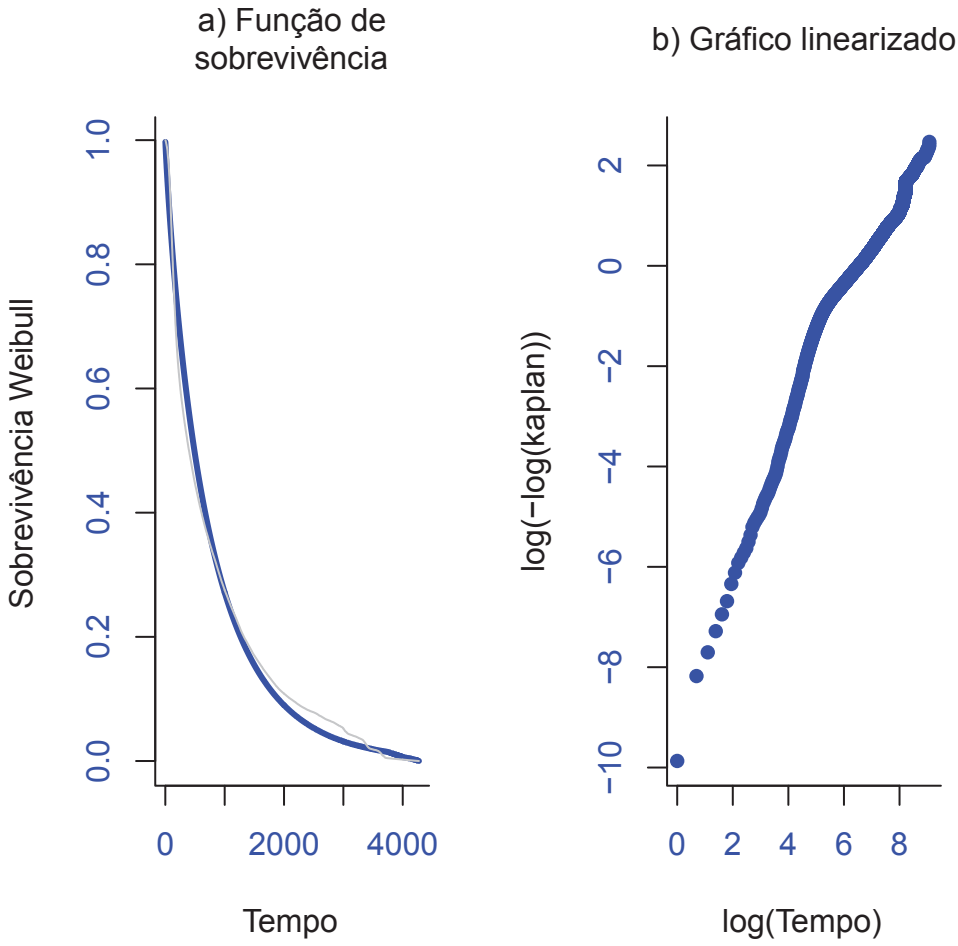


Fonte: Dados da pesquisa

O gráfico b) mostra a função linearizada. Se o modelo for apropriado, espera-se que o gráfico seja aproximadamente linear. Neste gráfico, o eixo x apresenta os tempos e o y, o logaritmo dos valores estimados por Kaplan Meyer.

Como no caso da exponencial, a Figura 28 a) compara a distribuição teórica de Weibull com a função de sobrevivência dos dados e o gráfico b) mostra a função linearizada. No gráfico a), o afastamento da função teórica, representada pela linha azul, começa a partir do tempo 2.000. O gráfico b) mostra, no eixo x, os logaritmos dos tempos e no eixo y, log(-log(Kaplan Meyer)), conforme exposto na Tabela 2. Neste gráfico, o esperado também é uma linha reta.

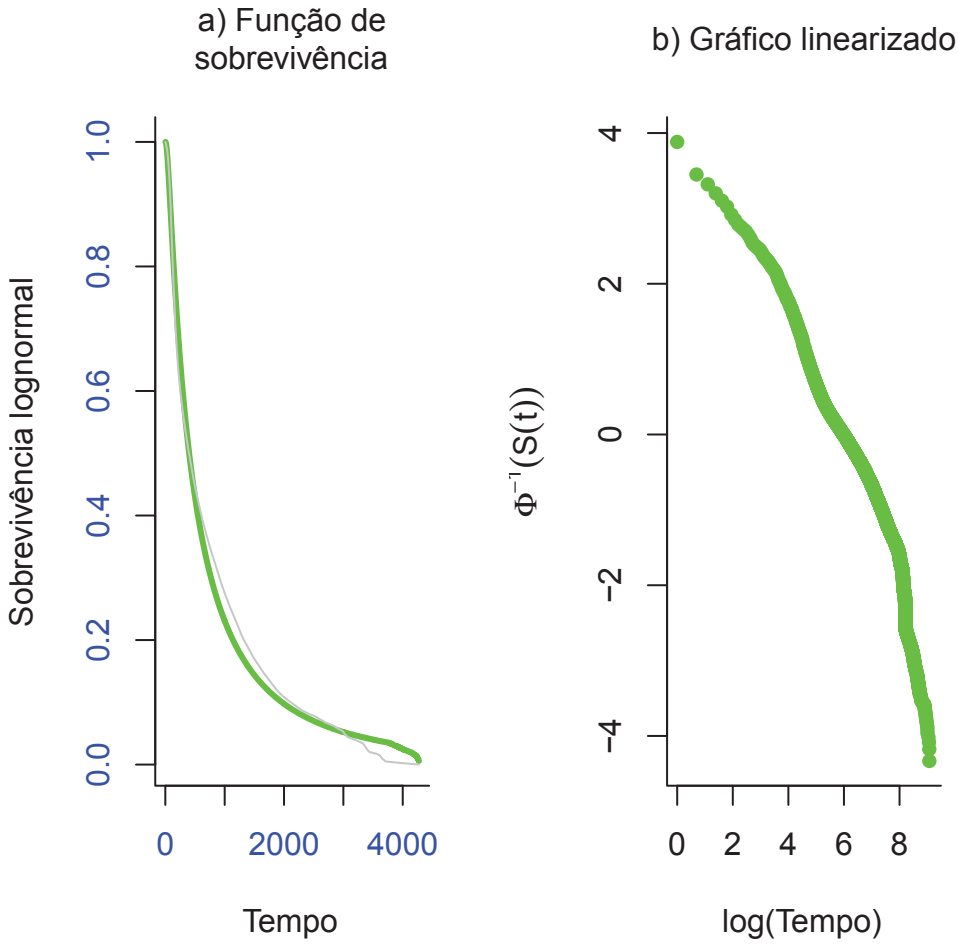
Figura 28 – Comparação com a distribuição Weibull



Fonte: Dados da pesquisa

A última distribuição comparada é a lognormal, conforme a Figura 29. O gráfico a) compara a distribuição teórica lognormal, em verde, com a função de sobrevivência estimada por Kaplan Meyer, em cinza. As discrepâncias em relação a uma lognormal começam a partir do tempo de 3.000 dias.

Figura 29 – Comparação com a distribuição lognormal



Fonte: Dados da pesquisa

A forma esperada do gráfico b), linearizado, é uma reta. Embora o ajuste não seja perfeito em nenhuma das distribuições, o da lognormal apresentou as menores discrepâncias entre as três distribuições, tanto para a comparação da distribuição de Kaplan Meyer quanto para o gráfico linearizado, avaliada apenas qualitativamente.

Contudo, testes para verificação do ajuste, como o de modelos encaixados, baseado na verossimilhança do modelo proposto e na verossimilhança de uma gamma generalizada, foram aplicados às três distribuições. Os p.values dos três testes resultaram em 0,000, embora a estatística do teste da lognormal tenha sido inferior: 12374, 8728 e 460, respectivamente, para as distribuições exponencial, Weibull e lognormal. A hipótese nula do teste é a adequação de ajuste do modelo.

Também foram realizados os testes de aderência do tempo de atravessamento às distribuições Weibull, exponencial e lognormal mediante a realização do teste de Kolmogorov-Smirnov para uma amostra, com o teste das seguintes hipóteses:

- Exponencial. Foi verificada a hipótese nula H0: a distribuição é exponencial, com parâmetro = 0,00126. A estatística do teste, calculada com base na distância vertical entre os valores da distribuição acumulada presumida para os dados (exponencial) e os valores da distribuição empírica, foi de 0,12679 (p.value=0,00);
- Weibull. A hipótese nula testada foi H0: a variável tempo segue uma distribuição Weibull com parâmetros de 0,8856 e 741.0641, para respectivamente forma e escala. A estatística do teste resultou em 0,080338 (p.value=0,00);
- Lognormal. Foi testada a hipótese nula H0: a variável tempo segue uma distribuição lognormal com parâmetros de 5,9954 e 1,2394, para respectivamente μ e σ. O valor da estatística calculada foi de 0,05522 (p.value=0,00).

Embora o p.value tenha sido baixo para as três distribuições, o que era esperado pelo elevado tamanho da amostra, a estatística do teste para a distribuição lognormal foi menor, indicando melhor ajuste. Para comparação complementar, a Tabela 7 apresenta os percentis das três distribuições ajustadas aos dados e o observado.

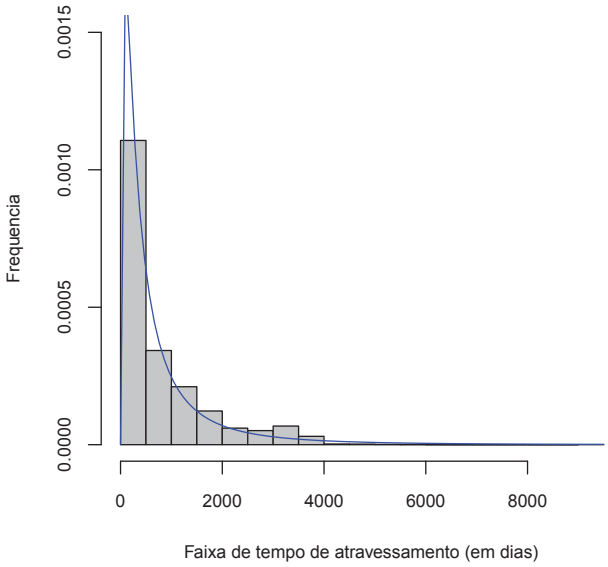
Tabela 7 – Percentis do tempo: observado e ajustado

Percentil	Observado	Distribuição teórica		
		Exponencial	Weibull	Lognormal
0,01	25	7,9	4,1	22,5
0,05	63	40,6	25,9	52,3
0,1	89	83,3	58,4	82,0
0,25	149	227,5	181,5	174,1
0,5	400	548,1	489,9	401,6
0,75	1103	1.096,2	1.071,6	926,4
0,9	2112	1.820,7	1.900,5	1.965,9
0,95	3021	2.368,8	2.558,2	3.084,0
0,99	3670	3.641,5	4.157,3	7.176,9

Fonte: Dados da pesquisa

Ainda que os percentis das distribuições tenham se apresentado discrepantes em relação às distribuições teóricas, os da distribuição lognormal apresentaram-se mais próximos ao observado, sendo que as diferenças começam a partir do 95^o percentil. O gráfico da Figura 30 mostra a densidade de uma distribuição lognormal com parâmetros $\mu = 5,995$ e $\sigma = 1,239$ sobreposta ao gráfico da função densidade dos dados. Estes ajustes foram obtidos com o pacote base do software R.

Figura 30 – Histograma do tempo de atravessamento sobreposto a distribuição lognormal



Fonte: Dados da pesquisa

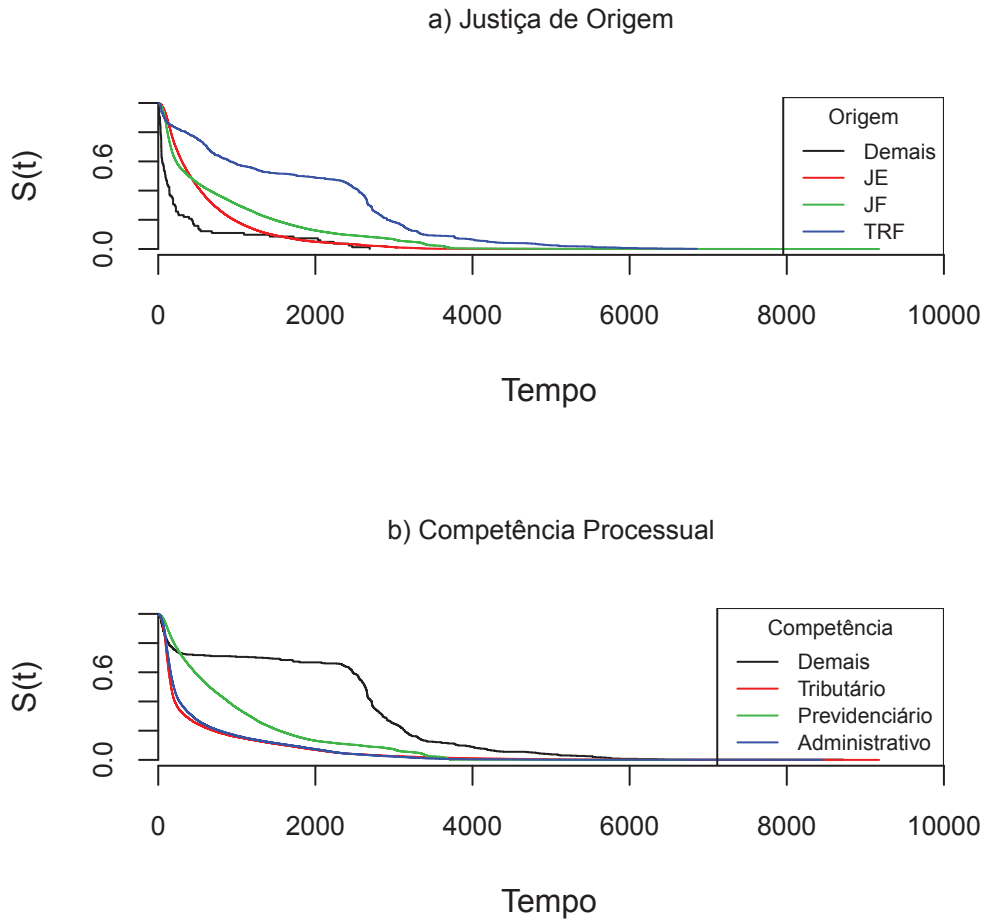
Desta forma, as hipóteses de ajustes foram rejeitadas, ao nível de significância de 0,01. Porém, como os testes estatísticos formais para falta de ajuste tendem a ter baixa potência para amostras pequenas ou rejeitar um dado modelo para amostras grandes, a decisão de ajuste dos modelos foi baseada nas verificações gráficas apenas. Assim, a distribuição lognormal foi usada para o ajuste do modelo.

O comportamento do risco observado, conforme a Figura 25, é semelhante ao de uma lognormal até o tempo de 6.600 dias: a função taxa de risco de uma lognormal caracteriza-se por ser inicialmente crescente e, a partir do ponto de máximo, decrescer. Desta forma, o ajuste a uma lognormal até um tempo corrobora com o encontrado na análise descritiva e no ajuste dos dados.

Como características da lognormal podem ser citadas: i) assume apenas valores positivos, ii) possui assimetria à direita, e como consequência, média superior à mediana. O caso estudado apresenta estas características: assimetria à direita e valores positivos apenas.

As funções de sobrevivência das covariáveis estão no Apêndice D. As diferenças entre as categorias das variáveis quanto ao tempo de atravessamento, podem ser comprovadas pelos gráficos de sobrevivência estimada. Por exemplo, quanto à justiça de origem, conforme o gráfico a) da Figura 31, a curva dos processos oriundos da Justiça Federal toca e ultrapassa a curva do tempo dos processos oriundos da Justiça Estadual em aproximadamente 400 dias, indicando que, para tempos menores que este valor, os processos provenientes da Justiça Federal tem maior probabilidade de baixa que os da Justiça Estadual. Mas, à medida que o tempo passa, a probabilidade dos processos vindos da Justiça Estadual baixar é maior.

Figura 31 – Justiça de origem e Competência

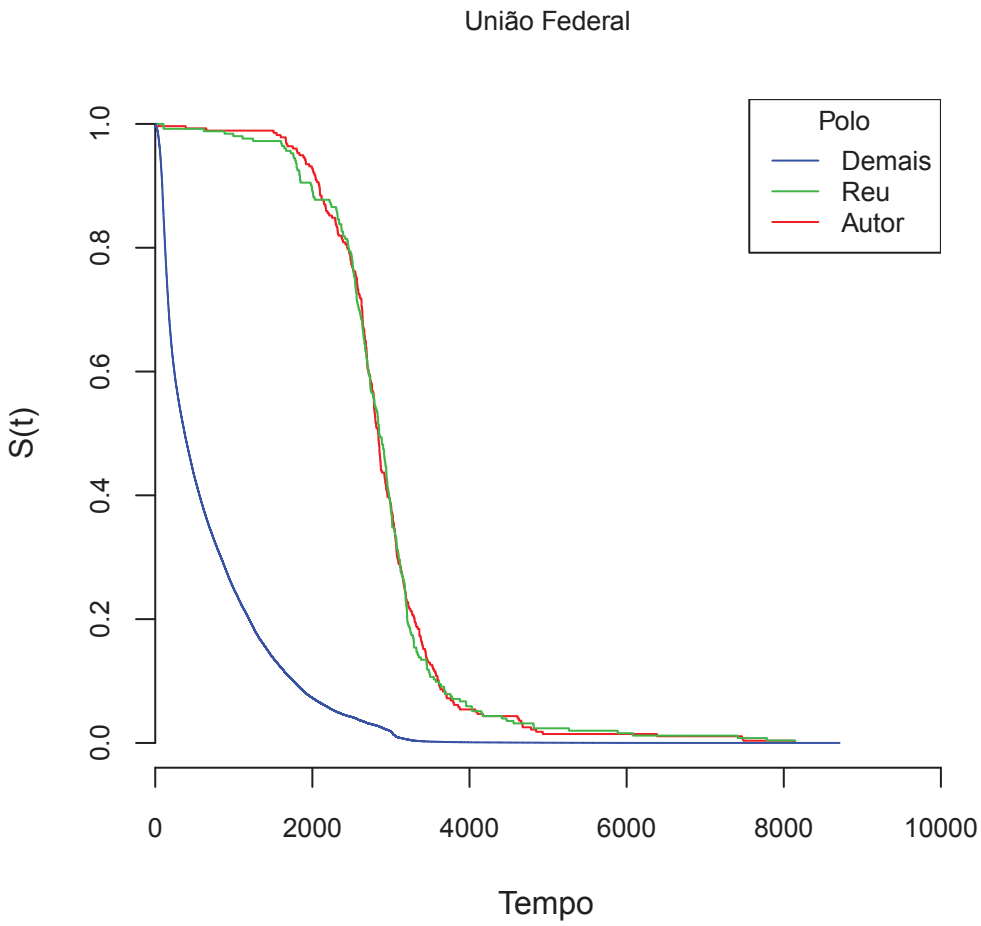


Fonte: Dados da pesquisa

No gráfico b) da Figura 31, as curvas das competências administrativo e tributário estão muito próximas, embora a curva de Tributário apresente probabilidade menor até aproximadamente 2.500 dias. A partir deste valor, a probabilidade dos processos administrativos baixar é maior, conforme a Figura 31.

Quanto às entidades, destaque-se a entidade União Federal, conforme mostrado no gráfico da Figura 32. A curva da União Federal ré ou autora está a frente das demais entidades (todas as demais), indicando o maior tempo de atravessamento destes processos.

Figura 32 – Função de sobrevivência União Federal



Fonte: Dados da pesquisa

5.2.1 Escolha das covariáveis

Inicialmente, foram realizados ajustes univariados com as covariáveis para verificar a influência das mesmas no tempo de atravessamento. Assim, foram testadas 102 hipóteses sob o modelo $Y = \mu + \gamma' Z + \sigma W$, onde Z_i é uma das 102 covariáveis, μ é o intercepto, γt é o coeficiente da variável Z e W, os resíduos. Foram testadas as hipóteses nulas H0: O coeficiente é igual a zero. Ao nível de significância de 0,10; as covariáveis unidade da federação Paraná, as entidades MPF (réu), ELETROBRAS (autor), EMGEA (autor), EMGEA (réu), INCRA (réu), IBAMA (réu), INMETRO (réu), Universidade (autor), a classe C11011 (Ação Rescisória - Seção), e o assunto N040101 (Expurgos Inflacionários) não foram consideradas significativas. Porém, estas variáveis foram mantidas durante a aplicação do método *stepwise* para escolha das covariáveis. No ajuste do modelo final, daquelas covariáveis com p.value superior a 0,10 no ajuste univariado, todas foram consideradas significantes, ao nível de significância de 0,05.

No ajuste de modelos de análise de sobrevivência não são usadas todas as covariáveis.

Dada uma certa variável, a quantidade de covariáveis é igual ao número de categorias $(p) - 1$. Por exemplo, unidade da federação possui três categorias: PR, RS e SC, mas para o ajuste, duas são suficientes. Assim, as categorias Demais foram suprimidas de classe, assunto, justiça de origem, competência, demais entidades ré e autora. O meio processual físico e a covariável unidade da federação Paraná também foram suprimidos. Como foi reportado no Capítulo 4, a variável gabinete não foi incluída nos ajustes.

As variáveis foram selecionadas utilizando a função *step* do pacote *stats*, que seleciona as variáveis através do critério AIC - Critério de Informação de Akaike, baseado no logaritmo da função de verossimilhança. Foi escolhida a direção *both* (*forward* e *backward*). Além da exclusão das variáveis sugeridas pela rotina, a variável C10855 (Reexame Necessário Cível) foi retirada do modelo por apresentar p.value superior a 0,05. Isto é, após o ajuste do modelo, os coeficientes são testados pela hipótese nula H0: o valor do coeficiente é igual a zero. Considerando distribuição normal, ao nível de significância de 0,05, não há evidências de que o coeficiente de Reexame Necessário Cível seja diferente de zero. As covariáveis finais do modelo foram:

- Meio processual. Ao meio Eletrônico foi atribuído o valor 1;
- Unidade da Federação: Rio Grande do Sul e Santa Catarina. O Paraná não consta pois o número de variáveis é igual ao número de categorias menos 1;
- Competência: Administrativo, Previdenciário, Tributário;
- Entidades: ANTT (réu e autor), CEF (réu e autor), Conselhos Regionais e Federais (réu e autor), Estado (réu e autor), INSS (réu e autor), MPF (réu e autor), Município (autor), AGU (autor e réu), Fazenda Nacional (autor e réu), União Federal (autor e réu), Universidades Federais (autor e réu), FNDE (autor e réu), IBAMA (autor), EMGEA (autor e réu), INMETRO (autor), Banco (réu), DNIT (autor e réu), ELETROBRAS (autor e réu), INCRA (autor), Instituto Federal (autor e réu);
- Classes: C10828 (Apelação Cível), C10992 (Apelação / Reexame Necessário), C10822 (Agravo de Instrumento), C10855 (Reexame Necessário Cível), C11011 (Ação Rescisória - Seção), C10943 (Ação Rescisória), C10977 (Embargos Infringentes), C11009 (Mandado de Segurança - Turma), C40012 (Pedido de Efeito Suspensivo á Apelação - Turma);
- Assuntos: N040119 (Aposentadoria por Tempo de Contribuição (Art. 55/6)), N0312 (Dívida Ativa), N040310 (Renúncia ao benefício), N040105 (Auxílio-Doença Previdenciário), N040101 (Aposentadoria por Invalidez (Art. 42/7)), N040104 (Aposentadoria Especial (Art. 57/8)), N04010202 (Aposentadoria por Idade - Rural (art. 48/51)), N040118 (Aposentadoria por Tempo de Serviço (Art. 52/4)), N040102 (Aposentadoria por Idade (Art. 48/51)), N04020113 (IRSM de Fevereiro de 1994(39,67%)), N06040101 (Expurgos Inflacionários / Planos Econômicos), N01040411 (Fornecimento de Medicamentos), N040103 (Aposentadoria por Tempo de Serviço (Art. 52/6) e/ou Tempo de Contribuição), N040203 (Reajustes e Revisões Específicos), N01031001 (Multas e demais Sanções), N010808

(Seguro-desemprego), N011501 (Multas e demais Sanções), N04020116 (Alteração do coeficiente de cálculo do benefício), N040501 (Averbação/Cômputo/Conversão de tempo de serviço especial), N040201 (RMI - Renda Mensal Inicial), N040107 (Salário- Maternidade (Art. 71/73)), N0219033205 (Seguro), N0402 (RMI - Renda Mensal Inicial, Reajustes e Revisões Específicas), N040113 (Benefício Assistencial (Art. 203,V CF/88)), N04020108 (Limitação do salário-de-benefício e da renda mensal inicial), N0401 (Benefícios em Espécie), N040404 (Concessão), N04011301 (Pessoas com deficiência), N02190405 (Cédula de crédito rural), N0115 (Dívida Ativa não-tributária), N040502 (Averbação/Cômputo de tempo de serviço de segurado especial (regime de economia familiar)), N030201 (IRPF/Imposto de Renda de Pessoa Física), N03040202 (Cofins), N040111 (Auxílio-Acidente (Art. 86)), N02190338 (Contratos Bancários).

Após a escolha das covariáveis, foi ajustado um modelo de regressão paramétrico de tempo de vida acelerado, utilizando as covariáveis mencionadas e a distribuição lognormal. Uma alternativa a este modelo é o modelo semi-paramétrico de Cox, que apresenta alta flexibilidade, é amplamente usado em análise de sobrevivência e pressupõe proporcionalidade das taxas de falhas.

Porém, o modelo de Cox não foi aplicado porque não foi comprovada, pela análise gráfica das funções estimadas de Kaplan Meyer para as covariáveis, a suposição de proporcionalidade das taxas de falhas. Por exemplo, para a variável justiça de origem, na Figura 31, as curvas da Justiça Federal e Estadual cortam-se, violando, desta forma, esta pressuposição . Os parâmetros do modelo ajustado estão na Tabela 8. O modelo ajustado foi: $Y = \mu + \gamma' Z + \sigma W$, onde $\gamma' = (\gamma_1, ..., \gamma_p)$ é o vetor dos coeficientes de regressão e W é a distribuição do erro, onde supõe-se que o tempo siga uma distribuição lognormal.

Tabela 8 – Estimativas dos parâmetros do modelo de regressão ajustado aos dados de treinamento

Variável	Valor	Erro padrão	z	p.valor
Intercepto	6,15	0,0296	208,09	0,0000
Eletrônico	-1,24	0,0104	-120,12	0,0000
SC	-0,18	0,0087	-20,98	0,0000
RS	-0,20	0,0076	-25,86	0,0000
Administrativo	0,32	0,0212	14,98	0,0000
Previdenciário	0,26	0,0349	7,46	0,0000
Justiça Estadual	-0,80	0,0113	-70,42	0,0000
TRF	0,39	0,0696	5,58	0,0000
CEF Autor	-0,38	0,0234	-16,33	0,0000
CEF Ré	-0,28	0,0195	-14,63	0,0000
Conselho Autor	-0,22	0,0225	-9,96	0,0000

Continua...

Tabela 8 – Estimativas dos parâmetros do modelo de regressão ajustado aos dados de treinamento ...continuação

Variável	Valor	Erro padrão	z	p.valor
Estado Autor	0,24	0,0472	5,14	0,0000
Estado Réu	-0,17	0,0431	-4,04	0,0001
INSS Autor	0,20	0,0288	6,99	0,0000
INSS Réu	0,40	0,0285	14,09	0,0000
MPF Autor	0,49	0,0534	9,14	0,0000
MPF Réu	0,33	0,0407	8,05	0,0000
Município Autor	0,32	0,0495	6,40	0,0000
Município Réu	0,17	0,0289	5,98	0,0000
AGU Autor	0,07	0,0220	2,96	0,0031
AGU Ré	0,11	0,0205	5,33	0,0000
Fazenda Nacional Autor	0,17	0,0239	7,15	0,0000
Fazenda Nacional Ré	0,19	0,0228	8,56	0,0000
União Federal Autor	0,77	0,0675	11,35	0,0000
União Federal Ré	0,89	0,0684	13,06	0,0000
Universidade Autor	0,09	0,0381	2,30	0,0217
FNDE Réu	-0,25	0,0768	-3,29	0,0010
IBAMA Autor	0,35	0,0468	7,39	0,0000
IBAMA Réu	0,17	0,0511	3,40	0,0007
EMGEA Autor	0,46	0,0963	4,75	0,0000
EMGEA Réu	0,36	0,0804	4,52	0,0000
INMETRO Autor	-0,18	0,0449	-3,90	0,0001
INMETRO Réu	0,41	0,0832	4,98	0,0000
Banco Autor	0,16	0,0547	2,89	0,0039
DNIT Autor	-0,16	0,0459	-3,45	0,0006
DNIT Réu	-0,22	0,0516	-4,21	0,0000
ELETROBRAS Autor	0,84	0,0598	14,12	0,0000
ELETROBRAS Ré	0,78	0,0919	8,49	0,0000
INCRA Autor	0,41	0,0872	4,75	0,0000
INCRA Réu	0,26	0,0958	2,72	0,0065
Institutos Federais Autor	-0,40	0,0653	-6,09	0,0000
Apel. Cível	0,86	0,0154	56,05	0,0000
Agr. de instrumento	0,27	0,0161	16,80	0,0000
Apel/Reexame necessário	1,15	0,0171	67,40	0,0000
Embargos infrintentes	1,91	0,0359	53,24	0,0000
Ação rescisória - seção	0,46	0,0777	5,90	0,0000

Continua...

Tabela 8 – Estimativas dos parâmetros do modelo de regressão ajustado aos dados de treinamento ...continuação

Variável	Valor	Erro padrão	z	p.valor
Pedido de Efeito Suspensivo	-0,64	0,0925	-6,92	0,0000
Aposentadoria por Tempo de Contribuição	0,15	0,0151	10,03	0,0000
Dívida Ativa	-0,09	0,0159	-5,93	0,0000
Renúncia ao benefício	0,30	0,0161	18,45	0,0000
Auxílio-Doença Previdenciário	-0,37	0,0171	-21,82	0,0000
Aposentadoria por Invalidez	-0,31	0,0190	-16,35	0,0000
Aposentadoria Especial	-0,10	0,0187	-5,47	0,0000
Aposentadoria por Idade - Rural	0,08	0,0196	4,13	0,0000
Aposentadoria por Tempo de Serviço	0,70	0,0233	29,83	0,0000
Aposentadoria por Idade	0,08	0,0256	3,14	0,0017
IRSM de Fevereiro de 1994	1,14	0,0530	21,43	0,0000
Expurgos Inflacionários / Planos Econômicos	-0,36	0,0264	-13,81	0,0000
Fornecimento de Medicamentos	-0,40	0,0325	-12,25	0,0000
Aposentadoria por Tempo de Serviço	1,28	0,0459	27,93	0,0000
Reajustes e Revisões Específicos	-0,16	0,0321	-4,95	0,0000
Multas e demais Sanções	-0,26	0,0342	-7,68	0,0000
Seguro-desemprego	-0,81	0,0341	-23,89	0,0000
Multas e demais Sanções	-0,26	0,0340	-7,68	0,0000
Alteração do coeficiente de cálculo do benefício	0,68	0,0351	19,34	0,0000
Averbação, Cômputo,Conversão de tempo de serviço especial	0,27	0,0366	7,40	0,0000
RMI - Renda Mensal Inicial	-0,23	0,0363	-6,24	0,0000
Salário-Maternidade	-0,21	0,0370	-5,75	0,0000
Seguro	-0,19	0,0394	-4,77	0,0000
RMI - Reajustes e Revisões	0,74	0,0379	19,62	0,0000
Benefício Assistencial	-0,20	0,0382	-5,21	0,0000
Limitação do salário-de-benefício e da RMI	-0,71	0,0408	-17,31	0,0000
Benefícios em Espécie	0,74	0,0412	17,90	0,0000
Concessão	0,80	0,0415	19,26	0,0000
Pessoas com deficiência	-0,76	0,0427	-17,89	0,0000
Cédula de crédito rural	-0,37	0,0432	-8,51	0,0000
Dívida Ativa não-tributária	-0,15	0,0430	-3,52	0,0004
Averbação/Cômputo de tempo de serviço de segurado especial	0,19	0,0445	4,24	0,0000

Continua...

Tabela 8 – Estimativas dos parâmetros do modelo de regressão ajustado aos dados de treinamento ...continuação

Variável	Valor	Erro padrão	z	p.valor
IRPF/Imposto de Renda de Pessoa Física	-0,12	0,0441	-2,62	0,0088
Cofins	-0,38	0,0461	-8,29	0,0000
Auxílio-Acidente (Art. 86)	-0,46	0,0452	-10,13	0,0000
Contratos Bancários	-0,29	0,0458	-6,43	0,0000

Fonte: Dados da pesquisa

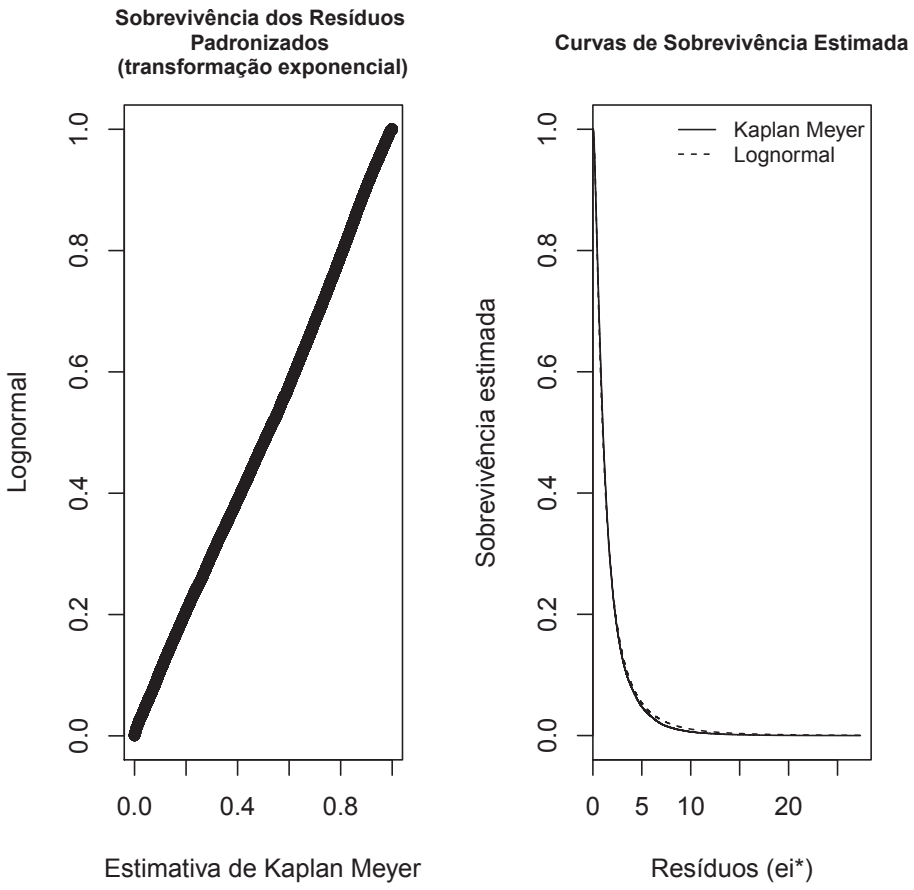
A interpretação dos coeficientes da tabela pode ser realizada a partir da exponencial dos coeficientes estimados no modelo. Assim, é obtida a razão dos tempos medianos de sobrevivência, mantendo as demais variáveis constantes. Para o modelo em questão, seguem exemplos de interpretação dos coeficientes ajustados:

- O tempo mediano de um processo eletrônico é 0,3 vezes o de processos físicos;
- O tempo mediano de um processo previdenciário é 1,3 vezes o de um processo não previdencário;
- O tempo mediano de um processo oriundo da Justiça Estadual é 0,5 vezes menor que um processo oriundo de outro tipo de justiça;
- Um processo oriundo da classe C10977 (embargos infringentes) possui tempo mediano 6,7 vezes maior que um processo de outra classe;
- O tempo mediano em que um processo da União Federal ré é 2,54 vezes maior que o das demais entidades ré.

5.2.2 Adequação do modelo

A adequacidade do modelo foi realizada de forma gráfica, mediante análise dos resíduos padronizados e de Cox-Snell. O gráfico a) da Figura 33 mostra as probabilidades de sobrevivência dos resíduos estimados por Kaplan Meyer e pelo modelo lognormal. Foi aplicada a transformação exponencial para produzir resíduos de uma distribuição lognormal. Como pode ser observado, a distribuição dos resíduos aproxima-se de uma reta, indicando um bom ajuste para os dados.

Figura 33 – Sobrevivência dos resíduos

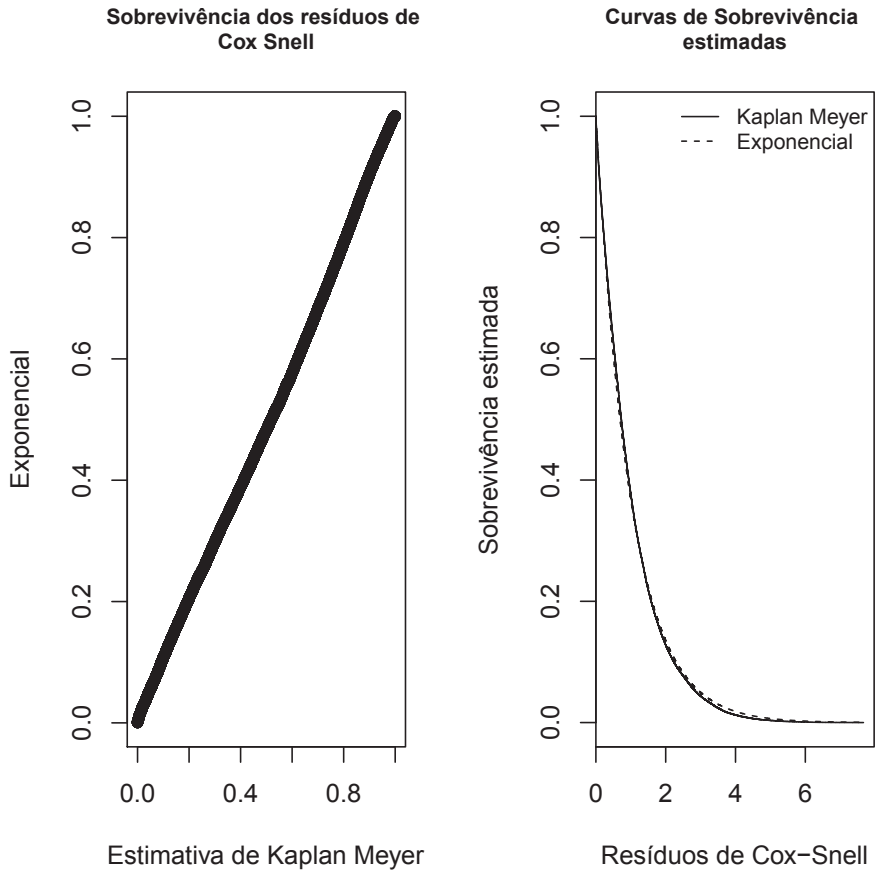


Fonte: Dados da pesquisa

O gráfico b) da Figura 33 mostra a função de sobrevivência dos resíduos pelo modelo lognormal, na linha pontilhada e a estimada por Kaplan Meyer para os resíduos na linha contínua. As linhas praticamente sobrespostas indicam, visualmente, um ajuste adequado.

Na sequência, foram analisados os resíduos de Cox-Snell, apresentados na Figura 34. O objetivo da análise destes resíduos é examinar o ajuste global do modelo e, se o modelo for adequado, estes resíduos devem seguir uma distribuição exponencial.

Figura 34 – Resíduos de Cox-Snell



Fonte: Dados da pesquisa

O gráfico a) na Figura 34 mostra a sobrevivência dos resíduos estimada por Kaplan Meyer (eixo x) e a sobreviência estimada pela exponencial no eixo y. Visualmente, o modelo lognormal parece ajustar-se bem aos dados, uma vez que aproxima-se de uma reta. O gráfico b) compara a sobrevivência dos resíduos de Cox Snell estimada por Kaplan Meyer e pela exponencial. Visualmente, há praticamente uma sobreposição de curvas, indicando o bom ajuste.

Confirmada graficamente a adequacidade do modelo ajustado, ele então foi aplicado ao conjunto de teste para obter valores preditos. Foi calculada a raiz quadrada média do erro de predição, RMSE, para mostrar o quanto as predições estão próximas aos valores observados. O RMSE calculado foi de 615,15 dias, tendo como base o conjunto dos dados de teste.

Os objetivos do ajuste de um modelo de análise de sobrevivência foram compreender a influência das covariáveis no tempo e obter predições para comparação com os demais modelos de regressão estimados. Para ilustrar o quanto a inserção das covariáveis no modelo diminuiu a medida de erro, foi ajustado um modelo sem a utilização de qualquer covariável, utilizando os dados de treinamento. O modelo foi então aplicado ao conjunto de dados de teste. O RMSE resultante foi de 819,6, superior ao encontrado com a utilização das covariáveis (615,15),

indicando que a inserção das covariáveis melhorou a previsão do tempo. As seções seguintes continuam com o ajuste de modelos de regressão de redes neurais e máquinas de vetor suporte.

5.3 Regressão por Redes Neurais

Com o objetivo de identificar um modelo com melhor ajuste, isto é, com RMSE menor, foi ajustado um modelo de regressão por redes neurais. As subseções seguintes descrevem as etapas seguidas para a obtenção deste modelo.

Como foi mencionado no Capítulo 4, para encontrar os parâmetros dos modelos de regressão, tanto utilizando redes neurais como máquina de vetor suporte, foi usada uma amostra de dados a fim de tornar o procedimento mais rápido. A partir da definição dos parâmetros obtidos com uma amostra de 20.000 casos, o modelo foi ajustado ao conjunto de dados de treinamento.

5.3.1 Transformação da variável resposta

O primeiro passo foi transformar a variável resposta. A transformação usada teve o objetivo de obter a variável como um valor entre 0 e 1. A transformação usada foi a seguinte.

$$y = \frac{Tempo}{\max(Tempo)} \tag{68}$$

As covariáveis descritas na seção 5.1.1 foram mantidas para o ajuste do modelo. Estas covariáveis apresentaram valores 0 e 1: por exemplo, a variável competência do processo apresentou 4 categorias, resultando, portanto, em quatro covariáveis: previdenciário, administrativo, tributário e demais. Por exemplo, se o processo for previdenciário, as covariáveis previdenciário, tributário, administrativo e demais competências assumem, respectivamente, valores 1, 0, 0, 0. Esta categorização foi a mesma para todos os quatro modelos ajustados.

5.3.2 Escolha do algoritmo

Após a definição da variável resposta, foram comparados o desempenho dos algoritmos retropropagação, retropropagação resiliente com e sem retrocesso de peso, através da medida RMSE. Para tanto, foi usado o pacote *neuralnet* mediante a comparação de duas camadas de neurônios ocultos, variando de 1 a 12, para os três algoritmos mencionados. O algoritmo retropropagação forneceu os menores erros, geralmente, inferior a 800, embora nem sempre tenha ocorrido a convergência devido ao grande número de covariáveis. Então, o restante da análise foi realizada com o pacote *deepnet*, conforme referido no Capítulo 4, com o algoritmo retropropagação.

O modelo foi aplicado ao conjunto de teste para a obtenção de valores preditos. Antes de obter o RMSE, os valores preditos foram transformados com o objetivo de obter o erro na mesma escala original, em dias. A transformação usada foi:

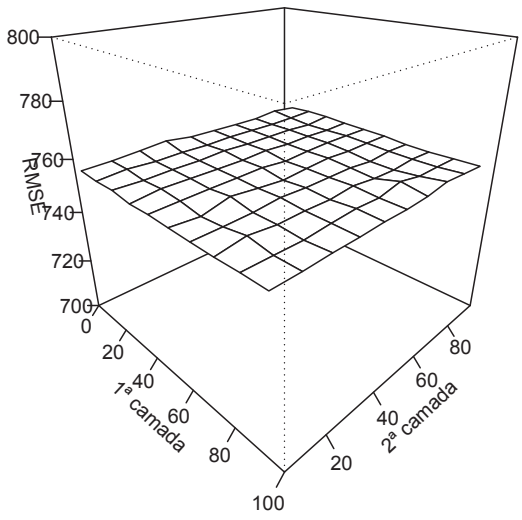
$$Tempo^{'} = y \times \max(Tempo) \tag{69}$$

onde y é o valor predito obtido e $Tempo^{'}$ é o valor previsto para o tempo em dias.

5.3.3 Escolha do número de neurônios e de camadas ocultas

Após, com o uso do pacote *deepnet*, foram testadas as combinações de taxa de aprendizado, quantidade de neurônios nas camadas ocultas. Especificamente, foram testadas duas camadas, com o número de neurônios variando de 1 a 100 em cada uma. A taxa de aprendizado assumiu os valores de 0,05; 0,1; 0,15; ..., 0,95. Nesta etapa, a quantidade de neurônios por camada retornou valores de RMSE próximos, como pode ser observado na Figura 35, que mostra o RMSE, na escala de dias, para duas camadas de neurônios ocultos e taxa de aprendizado de 0,9.

Figura 35 – RMSE de acordo com o número de neurônios nas camadas - algoritmo retropropagação



Fonte: Dados da pesquisa

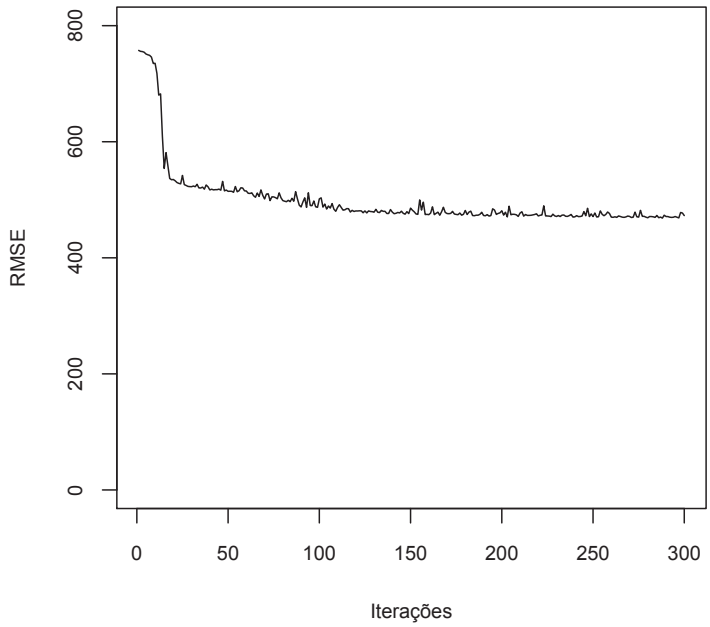
Assim, o modelo que retornou o menor RMSE foi o com duas camadas ocultas, com respectivamente 89 e 26 neurônios, taxa de aprendizado de 0,9. O RMSE para o modelo ajustado

com estes parâmetros foi de 750,2634.

5.3.4 Escolha do número de iterações

Após a escolha do número de camadas ocultas, de neurônios e taxa de aprendizado, foi testado o número de iterações, variando de 1 a 300, como critério de parada do algoritmo. O número de iterações escolhido foi 288, o qual resultou em RMSE de 470,4525, conforme resultado do gráfico da Figura 36

Figura 36 – RMSE de acordo com o número de iterações



Fonte: Dados da pesquisa

De acordo com o gráfico, após as 50 iterações, o valor de RMSE diminui de forma mais lenta e após a iteração 250, o RMSE estabiliza em torno de 470. As etapas para a obtenção destes parâmetros estão resumidas na Tabela 9.

Tabela 9 – RMSE de acordo com a etapa de obtenção do modelo - regressão

	Etapa	RMSE
Variável	$\frac{\max(Tempo)-Tempo}{\max(Tempo)-\min(Tempo)}$	737.4569
Resposta	$\frac{Tempo}{\max(Tempo)}$	737.0428
Algoritmo	retropropagação	750.2634
Número de neurônios e camadas ocultas	(89, 26)	750,2634
Taxa de aprendizado	0,9	750,2634
Número de iterações	288	470,4525

Fonte: Dados da pesquisa

5.3.5 Modelo ajustado e aplicação ao conjunto de teste

A seguir, com o objetivo de melhorar o RMSE, foram utilizados os pesos obtidos no ajuste anterior como pesos iniciais para a obtenção de um modelo final. Porém, esta inserção pouco diminuiu a medida, a qual passou para 470,2745. A Tabela 10 mostra os parâmetros usados para a obtenção deste modelo final.

Tabela 10 – Parâmetros para a estimação do modelo redes neurais

Parâmetro	Valor
Pacote	*deepnet*
Algoritmo	retropropagação
Número de camadas ocultas	2
Número de neurônios em cada camada	89 e 26
Taxa de aprendizado	0,9
Número de iterações	288
RMSE (dados de teste)	470,2745

Fonte: Dados da pesquisa

5.3.6 Avaliação do Modelo

A avaliação do modelo foi realizada através de *cross fold validation*, ou validação cruzada, com os procedimentos descritos no Capítulo 4. O conjunto de dados foi ordenado de forma aleatória e dividido em $k = 5$ partes iguais, neste caso, (D_i=25.548). Cada conjunto de tamanho igual ao conjunto de dados – D_i (102.192) foi considerado como um novo conjunto de treinamento e o seu complementar, como um novo conjunto de teste.

Para cada um dos cinco novos conjuntos de treinamento, o modelo foi ajustado, com os seguintes parâmetros: algoritmo retropropagação; duas camadas ocultas, com respectivamente, 89 e 26 neurônios em cada camada; taxa de aprendizado igual a 0,9; número de iterações igual a 288; e pesos iniciais iguais aos inseridos no modelo final.

O modelo obtido foi então aplicado ao novo conjunto de teste, o complementar ao conjunto de treinamento. Os valores preditos foram obtidos e transformados para a escala de dias. Após a obtenção dos 5 RMSEs, foram calculados a média e o desvio padrão, resultando em 468,9942 e 3,6483, para respectivamente, média e desvio padrão.

O RMSE obtido com a aplicação do modelo ajustado na fase de treinamento ao conjunto de teste, de 470,2745, está dentro do intervalo $\mu \pm \sigma$ (468,9942 $\pm$ 3,6483). Embora apresentando RMSE alto tanto no teste quanto na avaliação, o modelo foi aceito porque o desvio padrão é reduzido se comparado à média.

Em comparação ao modelo obtido através de análise de sobrevivência, os resultados da aplicação forneceram resultados melhores, embora o RMSE seja considerado alto. Por outro lado, redes neurais não fornecem informação sobre as covariáveis ou a comparação entre elas. Da mesma forma, redes neurais, à medida que não faz pressuposições sobre a distribuição das variáveis, dispensa comprovações como nas técnicas estatísticas formais.

5.4 Máquina de Vetor Suporte - Regressão

A seguir, a descrição dos procedimentos realizados para ajustar os modelos usando os algoritmos de máquina de vetor suporte para regressão e classificação. A análise foi realizada também em duas etapas: primeiro, o treinamento de um conjunto de dados para obter um modelo e segundo, a aplicação deste modelo ao conjunto de dados de teste.

A divisão do conjunto de dados em treinamento e teste, as covariáveis usadas bem como a amostra para a escolha dos parâmetros, tanto de regressão como classificação, foram os mesmos para o ajuste do modelo de redes neurais.

5.4.1 Escolha da variável resposta

O primeiro passo foi escolher a variável resposta. Assim, foram testadas três possibilidades: o tempo sem qualquer transformação, o logaritmo do tempo e a normalização. Os parâmetros usados foram default do pacote (tipo kernel, γ = 1/ dimensão dos dados, custo igual a 1). A transformação com menor RMSE foi a normalização (477,9230), obtida por meio de:

$$y = \frac{Tempo - \mu(Tempo)}{\sigma(Tempo)} \tag{70}$$

Os valores de RMSE para as demais propostas de variável resposta foram 477,9305 e 477,9238, respectivamente, para a variável tempo sem transformação e para o logaritmo do tempo.

Para o cálculo do RMSE, o modelo obtido com dados da amostra foi aplicado ao conjunto de teste para obtenção de valores preditos. Os valores preditos obtidos foram transformados para

obter o erro em dias. A transformação usada foi

$$Tempo^{'} = y \times \sigma(Tempo) + \mu(Tempo) \tag{71}$$

onde y é o valor predito e $Tempo^{'}$ é o valor previsto na escala de dias.

5.4.2 Escolha da função kernel, tipo de regressão e dos parâmetros custo, γ e υ

O prosseguimento da seleção dos parâmetros do modelo foi mediante a comparação das combinações de tipo de kernel (linear, polinomial, radial e sigmoide) e tipo de regressão (υ ou ε), resultando em oito modelos comparativos. O melhor resultado foi para a combinação de kernel radial e υ-regressão, com valor de RMSE de 472,2458. A Tabela 11 resume os passos realizados e apresenta o RMSE para cada etapa.

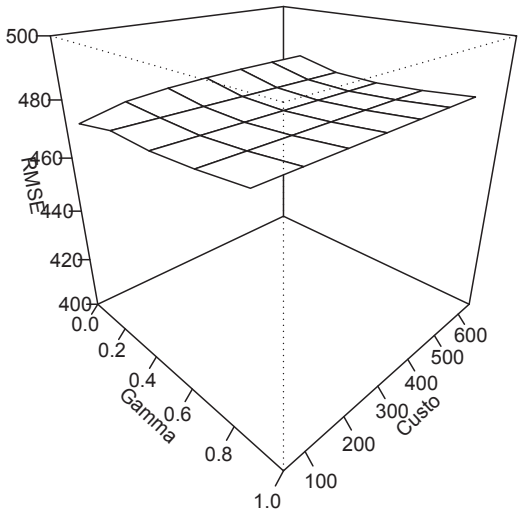
Tabela 11 – RMSE de acordo com a etapa de obtenção do modelo - regressão

Parâmetro	Etapa	RMSE
Variável Resposta	Tempo	477,9305
	$log(Tempo)$	477,9238
	$\frac{Tempo-\mu(Tempo)}{\sigma(Tempo)}$	477,9230
Tipo de regressão e Kernel	$\varepsilon - regressao$, linear	601,1557
	$\varepsilon - regressao$, polinomial	509,8597
	$\varepsilon - regressao$, radial	477,9230
	$\varepsilon - regressao$, sigmoid	477,4704
	$\upsilon - regressao$, linear	571,7511
	$\upsilon - regressao$, polinomial	503,9157
	$\upsilon - regressao$, radial	472,2458
	$\upsilon - regressao$, sigmoid	472,8805
Gamma, Custo e $\upsilon - regressao$	0,02; 100; 0,5	466,3132

Fonte: Dados da pesquisa

A partir deste resultado, foram testados diversos valores de γ, υ e custo: os valores foram $\gamma \in$ [0, 2; 0, 4; 0, 6, 0, 8; 1];custo $\in$ [100, 200, ..., 500] e υ-regressão $\in$ [0, 1; 0, 3, ..., 0, 9]. A combinação que forneceu o menor RMSE foi a de 0,2; 100 e 0,5; respectivamente, para γ, custo e υ. A Figura 37 mostra as combinações de γ e custo de acordo com o RMSE, para o valor de υ de 0,5.

Figura 37 – Máquina de vetor suporte - segundo valores de γ e custo



Fonte: Dados da pesquisa

Simultaneamente, foi usada a função tune, do pacote *e1071* do R, para a escolha dos melhores parâmetros para γ e custo. Os valores dos parâmetros dispostos na função foram: $\gamma \in$ (0; 0,1; 0,2; 0,3; 0; 4; 0; 5; 0,6; 0,7; 0,8; 0,9; 1) e custo $\in$ (4, 8, 16, 32, 64, 128, 256, 512). O valor retornado pela função foi a combinação de γ igual a 0,4 e custo igual a 512. O RMSE desta combinação de parâmetros obtida pela função tune foi igual a 476,4619.

5.4.3 Modelo ajustado e aplicação ao conjunto de teste

Após, ainda foram inseridos outros parâmetros para o modelo, na busca por melhores valores para o RMSE. A Tabela 12 apresenta os parâmetros para o modelo final.

Tabela 12 – Parâmetros usados para estimação do modelo Máquina de vetor suporte - regressão

Parâmetro	Valor
Pacote	*e1071*
Tipo de regressão	υ regressão
Kernel	Radial
função tune	476,4619
Gamma	0,02
Custo	100
Número de iterações	0,5
RMSE	466,3132

Fonte: Dados da pesquisa

O modelo final, obtido com os parâmetros indicados na Tabela 11, resultou em 45.571 vetores de suporte. O RMSE produzido, aplicando o modelo ao conjunto de dados de teste, foi de 466,3132.

5.4.4 Avaliação do modelo

Como no modelo ajustado em redes neurais, a avaliação foi realizada através de validação cruzada, com k = 5, utilizando-se da mesma ordenação e divisão do conjunto de dados. Após a avaliação, o RMSE médio para os 5 *folds* foi de 466,2799 e desvio padrão de 3,82. O RMSE calculado, usando o modelo obtido na fase de treinamento ao conjunto de dados do teste, de 466,3132 está dentro do intervalo $\mu \pm \sigma$ (466,2799 $\pm$ 3,82), obtido na avaliação. Como no modelo usado em redes neurais, o desvio padrão é baixo e o valor médio é próximo ao valor obtido na fase de teste.

Os valores de RMSE obtidos tanto na fase de avaliação, como o modelo final, estão próximos ao obtido com a utilização de redes neurais, e inferior ao obtido com a análise de sobrevivência. O objetivo de analisar com as duas técnicas, redes neurais e máquina de vetor suporte, é verificar qual delas fornece o menor RMSE, à medida que estas técnicas não avaliam sob o ponto de vista das covariáveis. Além disso, redes neurais e máquina de vetor suporte aprendem com os dados de treinamento, obtendo padrões que generalizam para dados futuros semelhantes.

Especificamente, máquinas de vetor suporte são úteis quando há muitas variáveis de entrada ou quando estas interagem com a variável de saída ou entre elas de forma não linear, tendo poucas suposições sobre as distribuições.

5.5 Máquina de vetor suporte - classificação

O termo predição refere-se tanto à previsão de rótulos de classe, utilizado pela classificação, quanto à previsão de valores de dados numéricos, como na regressão.

Em classificação, o objetivo é predizer uma classe de acordo com as diversas variáveis, neste caso, uma faixa de tempo de atravessamento. A técnica escolhida para a obtenção do modelo é máquinas de vetor suporte para classificação. Para os modelos, obtidos a partir de diferentes parâmetros, a medida escolhida foi a acurácia total, calculada a partir da comparação da tabela dos valores preditos com os observados.

Como nos ajustes de regressão por redes neurais e máquina de vetor suporte, a categorização da variável resposta e os parâmetros foram escolhidos através de modelos obtidos com dados da amostra. Estes modelos foram então aplicados ao conjunto de teste para obtenção de medida comparativa, neste caso, a acurácia total. A avaliação do modelo foi também procedida com o conjunto de dados ordenado e dividido da mesma forma.

5.5.1 Categorização para o tempo de atravessamento

Inicialmente, com o objetivo de escolher a faixa de tempo a ser utilizada, foram comparadas as seguintes faixas de tempo de atravessamento:

- Até 180 dias, de 180 a 365 dias, de 366 a 730 dias, de 731 a 1460 dias, de 1461 a 2920 dias, mais de 2920 dias. A acurácia total calculada foi de 48,3%;
- Até 365 dias; de 366 a 730 dias, de 731 a 1095 dias; de 1096 a 1460 dias; de 1461 a 1825 dias; de 1826 a 2190 dias, mais de 2190 dias. A acurácia total calculada foi de 56,0%;
- Até 365 dias, de 366 a 1095 dias, de 1096 a 1825 dias, de 1826 a 2920 dias, de 2921 dias a 4380 dias, mais de 4380 dias. A acurácia total é de 61,4%;
- Até 545 dias, de 546 a 1090 dias, de 1091 a 2180 dias, de 2181 2170 dias, mais de 2170 dias a 3270 dias, mais de 3270 dias. A acurácia total foi de 66,2%;
- Até 545 dias, de 546 a 1090 dias, de 1091 a 2180 dias, mais de 2181. A acurácia total foi de 66,59%;
- Até 730 dias, de 730 a 1460 dias, de 1461 a 2190 dias, de 2191 a 2920 dias, de 2921 a 3650 dias, de 3650 a 4380 dias, mais de 4380 dias; A acurácia total calculada foi de 71,7%.

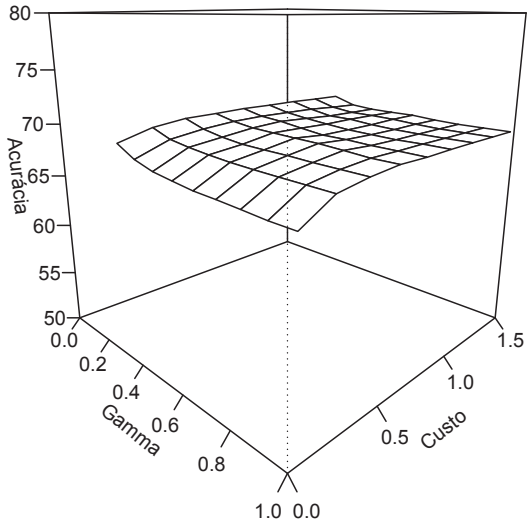
Foram testadas estas faixas de tempo por meio da biblioteca Libsvm e o modelo obtido foi aplicado ao conjunto de dados de teste. Os parâmetros usados foram default do pacote (tipo kernel, $\gamma = 1/$ dimensão dos dados, custo igual a 1). Após a obtenção das medidas de acurácia total, a categorização escolhida foi até 545 dias, de 546 a 1090 dias, de 1091 a 2180 dias, mais de 2181. Estas faixas de tempo representam, respectivamente, 57,3; 17,4; 15,8; 9,5% das observações.

Embora a acurácia tenha sido maior para a última proposta de classificação, esta foi rejeitada porque os intervalos de classe possuem amplitude de dois anos. Com esta classificação, o primeiro intervalo cobriria mais de 64,3% das observações. Além disso, o histograma, apresentado na Figura 23, possui uma forma assimétrica positiva, com a concentração de tempos na faixa de até 500 dias. Desta forma, é esperado que os intervalos de classe sejam desiguais e concentrados em valores de tempo menores em vez de intervalos de classe iguais. A seguir, com a categorização do tempo realizada, foram comparados quanto à acurácia total, as funções kernel linear, polinomial, radial e sigmoide. A função com a melhor acurácia foi a radial (66,6%). As demais acurácias totais resultaram em 65,5; 58,8 e 64,9%, respectivamente, para linear, polinomial e sigmoide. Inicialmente, os parâmetros são os default do pacote.

5.5.2 Escolha da função kernel dos parâmetros de γ e custo

Após, foram comparados diversos valores de custo e γ, utilizando o kernel radial. Os valores de γ e custo estão no gráfico da Figura 38. A combinação de valores com maior acurácia total para o modelo foram 0,1 e 1,2, respectivamente, para γ e custo. Para obter estes valores, foram sugeridos custo igual a 0,2; 0,4; ...; 1,4 e γ, 0,1; 0,2; ...; 1.

Figura 38 – Acurácia segundo valores de γ e Custo



Fonte: Dados da pesquisa

Como no ajuste do modelo de regressão com máquina de vetor suporte, simultaneamente, foi aplicada a função tune do pacote e1071. A função retornou, como valores de γ e custo, 0,2 e 1, respectivamente e acurácia total calculada de 68,7%. Para aplicação da função tune, os valores de γ e custo assumiram os valores de 0,2; 0,4;..., 1.

Então, os parâmetros do modelo estimado são γ igual a 0,1 e custo, 1,2. A Tabela 13 apresenta os parâmetros do modelo.

Tabela 13 – Parâmetros usados - Máquina de vetor suporte - classificação

Parâmetro	Valor
Pacote	e1071
Kernel	Radial
Gamma	0,1
Custo	1,2
Acurácia total	69,7%

Fonte: Dados da pesquisa

5.5.3 Modelo ajustado e aplicação ao conjunto de teste

O modelo com os parâmetros especificados foi ajustado ao conjunto de dados de treinamento, resultando em 48.277 vetores de suporte. O modelo foi então aplicado ao conjunto de dados de teste, obtendo-se valores preditos para as faixas de tempo. Foram então calculadas as medidas de acurácia, sensibilidade, especificidade, precisão e F1, conforme a Tabela 14.

Tabela 14 – Acurácia, sensibilidade, especificidade, precisão e F1 do modelo ajustado

Tempo em dias	Acurácia	Sensibilidade	Especificidade	Precisão	F1
Até 545	77,6	93,1	55,0	75,1	83,1
De 546 a 1090	80,6	20,9	94,1	44,1	28,3
De 1091 a 2180	83,3	37,9	92,1	48,0	42,4
Mais de 2181	97,7	72,7	99,5	90,1	80,5

Fonte: Dados da pesquisa

A tabela anterior foi obtida a partir de uma matriz de confusão para cada classe de tempo. Esta matriz foi gerada transformando a tabela de preditos versus observados dos dados de teste em tabelas 2×2, uma para cada classe, onde os rótulos são: positivo (da classe) e negativo (das outras classes).

A Tabela 14 mostrou acurácia maior para todas as faixas de tempo em comparação à acurácia total (69,7%). A sensibilidade, medida com o objetivo de verificar o que está corretamente classificado em relação a todos os itens da faixa de tempo, é maior para a categoria até 545 dias e menor para a de 546 a 1090 dias. O objetivo da medida especificidade é avaliar o que está corretamente classificado como de outras faixas de tempo em relação a todos os itens

das outras faixas. A especificidade é mais baixa para a faixa de até 545 dias, sendo esta medida superior a 90% para as demais faixas.

Assim, considerando como base a primeira classe, a com o maior número de observações, a acurácia é de 77,3%. A sensibilidade, de 93,6% para esta classe, indica que o teste está evitando falsos negativos (processos desta faixa classificados em outras). A especificidade indica que o teste classifica corretamente como de outras classes 53,4% dos processos que não são desta primeira faixa de tempo, sendo esta a medida com menor índice. A precisão desta primeira faixa de tempo, de 74,5%, indica o percentual de processos classificados corretamente em relação a todos os classificados desta faixa de tempo.

5.5.4 Avaliação do modelo

Após a obtenção das medidas descritas, o modelo foi testado, utilizando a técnica validação cruzada, a partir de $k = 5$, como em redes neurais e máquina de vetor suporte para regressão. A Tabela 15 mostra $\mu \pm \sigma$ das medidas calculadas a partir da avaliação e obtidas por meio das matrizes de confusão para cada uma das classes.

Tabela 15 – Acurácia, sensibilidade, especificidade, precisão e F1 na avaliação do modelo

Tempo em dias	Acurácia	Sensibilidade	Especificidade	Precisão	F1
Até 545	77,4 $\pm 0,2$	93,1 $\pm 0,3$	54,8$\pm 0,6$	74,7$\pm 0,4$	82,9 $\pm 0,2$
De 546 a 1090	80,6$\pm 0,3$	22,2$\pm 0,6$	93,7$\pm 0,4$	43,9$\pm 0,9$	29,5$\pm 0,4$
De 1091 a 2180	83,4$\pm 0,1$	37,0$\pm 0,7$	92,7$\pm 0,2$	50,2$\pm 1,1$	42,6$\pm 0,7$
Mais de 2181	97,8$\pm 0,1$	73,0$\pm 0,8$	99,4$\pm 0,1$	89,7$\pm 1,2$	80,5$\pm 0,5$

Fonte: Dados da pesquisa

5.6 Comparação entre os modelos

Todos os ajustes mostraram ou um erro grande, ou uma acurácia reduzida. A análise de sobrevivência mostrou o maior RMSE entre os modelos que analisaram a partir da variável resposta quantitativa. Os modelos que utilizaram redes neurais e máquina de vetor suporte mostraram medidas muito próximas, tanto na fase de teste quanto na de avaliação. A Tabela 16 resume os valores.

Tabela 16 – Comparação entre os modelos

Técnica	Variável resposta	Medida	Valor
Redes neurais (regressão)	Tempo em dias	RMSE	470,2745
Máquina de vetor suporte (regressão)	Tempo em dias	RMSE	466,3132
Análise de sobrevivência	Tempo em dias	RMSE	615,15
Máquina de vetor suporte (classificação)	Faixas de tempo	Acurácia	69,7%

Fonte: Dados da pesquisa

Da mesma forma, os três modelos avaliados, máquina de vetor suporte para regressão e classificação, redes neurais, a avaliação resultou em valores muito próximos ao obtido na fase de teste, além de apresentar, nos três casos, desvios reduzidos. Isto significa que o modelo treinado está se ajustando de forma estável aos dados. O modelo empregando análise de sobrevivência não foi avaliado.

De qualquer forma, como sugestão de um modelo para a predição, com o objetivo de informação às partes, é o que usou faixas de tempo em vez do valor numérico. O trabalho de Gruginskie e Vaccaro (2018), referente aos processos baixados em 2016, no TRF4, apresentou acurácia total semelhante, de 71,84%. Desta forma, embora tenha medidas reduzidas, a faixa de tempo é mais esclarecedora que o tempo em dias para quem recebe a informação sobre o tempo de atravessamento de um processo.

Por outro lado, a predição conforme o tempo mensurado em escala quantitativa se adequa para o estabelecimento de parâmetros e a verificação da conformidade com estes parâmetros. Neste caso, sugere-se ou a utilização do modelo obtido por redes neurais ou por máquina de vetor suporte porque o desempenho de ambos, medido por meio do RMSE, foi muito próximo.

De qualquer forma, as abordagens usadas são diferentes quanto aos pressupostos, em relação à variável resposta, além de outras particularidades. A seção 5.8 discute e compara os algoritmos e técnicas de análise, além de discutir sobre as variáveis dos modelos.

5.7 Características associadas ao Tempo processual

Para identificar as variáveis associadas a processos com maior tempo, foram realizadas análises de regras de associação. A partir da classificação dos processos em quatro faixas de tempo, e das variáveis classe, unidade da federação, assunto, meio, justiça de origem e competência do processo, foi realizada a análise com dois objetivos básicos:

- Geração de regras para identificar as características dos processos das faixas de tempo mais frequentes;
- Geração de regras para identificar as características dos processos mais demorados, isto é, os processos com maior tempo de atravessamento, mesmo que menos frequentes.

Para atingir estes objetivos foi utilizado o pacote *arules* em duas etapas, uma para cada objetivo. O conjunto de dados usados foi o total de processos, com a eliminação dos valores perdidos, totalizando 127.740 processos. Os parâmetros usados estão na Tabela 17.

Tabela 17 – Parâmetros usados na geração de regras de associação

Parâmetro	Processos mais frequentes	Processos mais lentos
Suporte mínimo	10%	0,5%
Suporte máximo	-	10%
Confiança	50%	40%
Mínimo de itens	2	2
Restrições	Lado direito = Faixa de Tempo	Lado direito = Faixa de Tempo excluindo a Faixa A

Fonte: Dados da pesquisa

A aplicação do algoritmo ao conjunto total de dados, na primeira etapa de geração de regras, resultou em 48, todas referentes à faixa de tempo de até 545 dias. Após a retirada das regras redundantes, restaram 24.

Estas regras foram úteis para identificar o perfil dos processos mais frequentes. Observam-se, a seguir, as regras 1, 3, 9 e 11. A regra número 1 foi constituída por processos eletrônicos oriundos do Rio Grande do Sul, da classe C10822 (agravo de instrumento), e com tempo de até 545 dias. Estes processos representam 11,4% do conjunto analisado, conforme a medida suporte. A confiança da regra fornece a seguinte informação: dos processos eletrônicos oriundos do Rio Grande do Sul, da classe C10822 (agravo de instrumento), 92,7% apresentaram tempo de atravessamento de até um ano e meio. Isto é, 7,3% destes processos apresentaram tempo superior a 545 dias.

Da mesma forma, os processos de competência administrativa, da classe C10822 (agravo de instrumento), oriundos da Justiça Federal e com tempo de atravessamento de até 545 dias representam 11,6% do conjunto analisado, de acordo com o suporte da regra. Dos processos administrativos, desta classe e oriundos da Justiça Federal, 88,9% apresentaram tempo de atravessamento de até 1 ano e meio, segundo a confiança da regra.

A regra número 9 refere-se a processos de competência tributária com tempo de atravessamento de até 545 dias. No conjunto analisado, eles representam 12,6% dos processos, segundo a medida de suporte. Do total de processos tributários, 77,3% apresentaram tempo de atravessamento de até um ano e meio. Isto é, 22,7% dos processos tributários apresentaram tempo de atravessamento superior a um ano e meio.

Já a regra número 11 refere-se a processos administrativos com tempo de atravessamento de até 545 dias. Esta regra representou 19,9% do total do conjunto do qual se extraíram as regras. Do total de processos administrativos, 74,7% apresentaram tempo de atravessamento de até 545 dias.

As regras número 19, 20, 21, 22 e 23 referem-se, respectivamente, a processos oriundos da Justiça Federal do Paraná, oriundos do Rio Grande do Sul, do Paraná, da Justiça Federal e de Santa Catarina. Apesar de todas estas regras apresentarem suporte superior a 10% e confiança

superior a 50%, a medida alavancagem está muito próxima de 1, especificamente entre 0,98 e 1,01. O valor da medida alavancagem indica falta de associação entre o lado direito (tempo de atravessamento) e esquerdo (características do processo). Ou seja, estas regras não indicam associação entre tempo e estas covariáveis. As regras estão na Tabela 18.

Tabela 18 – Regras para atingir o primeiro objetivo

Lado Esquerdo	Lado Direito (Faixa de Tempo)	Suporte	Confiança	Alavancagem
1 - Classe=C10822,UF=RS,Meio=Elet	l2=A	0,1142	0,9271	1,5743
2 - Classe=C10822,Meio=Elet	l2=A	0,2311	0,9215	1,5649
3 - Compe=Ad,Tipo=JF, Classe=C10822	l2=A	0,1159	0,8889	1,5096
4 - Compe=Ad,Classe=C10822	l2=A	0,1180	0,8884	1,5087
5 - Tipo=JF,Classe=C10822	l2=A	0,2024	0,8183	1,3897
6 - Compe=Tr,Meio=Elet	l2=A	0,1204	0,8115	1,3781
7 - Classe=C10822	l2=A	0,2332	0,8111	1,3774
8 - Compe=Ad,Meio=Elet	l2=A	0,1975	0,7790	1,3229
9 - Compe=Tr	l2=A	0,1258	0,7726	1,3120
10 - Compe=Ad,Tipo=JF	l2=A	0,1951	0,7481	1,2704
11 - Compe=Ad	l2=A	0,1988	0,7472	1,2689
12 - Meio=Elet,Assunto=Outro	l2=A	0,1896	0,7029	1,1937
13 - UF=PR,Meio=Elet	l2=A	0,1640	0,6657	1,1305
14 - Tipo=JF,Assunto=Outro	l2=A	0,1852	0,6590	1,1192
15 - Assunto=Outro	l2=A	0,1934	0,6550	1,1123
16 - UF=RS,Meio=Elet	l2=A	0,2285	0,6511	1,1057
17 - Meio=Elet	l2=A	0,5101	0,6459	1,0968
18 - Tipo=JE	l2=A	0,1694	0,6065	1,0300
19 - Tipo=JF,UF=PR	l2=A	0,1063	0,5964	1,0128
20 - UF=RS	l2=A	0,2810	0,5959	1,0120
21 - UF=PR	l2=A	0,1676	0,5868	0,9966
22 - Tipo=JF	l2=A	0,4184	0,5829	0,9899
23 - UF=SC	l2=A	0,1403	0,5776	0,9808
24 - Classe=C10828	l2=A	0,2392	0,5280	0,8967

Fonte: Dados da pesquisa

Legenda: A - até 545; B - de 546 a 1090; C - de 1091 a 2180; D - mais de 2180 dias C10822 – Agravo de instrumento, C10828 - Apelação cível Ad = Competência Administrativa, Tr =Competência Tributária

O prosseguimento da análise foi através da obtenção de regras para os processos com maior tempo de atravessamento. Para tanto, foi fixado, no lado direito da regra, as faixas de tempo maior que um ano e meio. Os resultados estão na Tabela 19.

Tabela 19 – Regras para atingir o segundo objetivo

Lado Esquerdo	Lado	Suporte	Confiança	Alavancagem
1 - Tipo=JF,Classe=C10828,Meio=0	D	0,0107	0,9920	15,8603
2 - Compe=Pr,Tipo=JF,UF=RS,Meio=0	D	0,0146	0,9909	15,8426
3 - Compe=Pr,Tipo=JF,Meio=0	D	0,0263	0,9759	15,6015
4 - Tipo=JF,UF=RS,Meio=0	D	0,0205	0,9632	15,3999
5 - Tipo=JF,Meio=0,Assunto=N040118	D	0,0060	0,9526	15,2290
6 - Tipo=JF,Meio=0	D	0,0399	0,9472	15,1429
7 - Classe=C10822,Meio=0, Assunto=N040118	D	0,0071	0,9210	14,7240
8 - Compe=Pr,Classe=C10822, UF=RS,Meio=0	D	0,0153	0,8933	14,2819
9 - Compe=Pr,Classe=C10822,Meio=0	D	0,0253	0,8750	13,9885
10 -Tipo=JF,Classe=C10822, Assunto=N040118	D	0,0052	0,8104	12,9571
11 - Classe=C10822,UF=RS,Meio=0	D	0,0171	0,8070	12,9013
12 - Classe=C10822,Assunto=N040118	D	0,0071	0,7981	12,7591
13 - Classe=C10822,Meio=0	D	0,0290	0,7890	12,6142
14 - Compe=Ad,Meio=0,Assunto=Outro	D	0,0080	0,7518	12,0200
15 - Compe=Ad,Meio=0	D	0,0093	0,7447	11,9054
16 - Meio=0,Assunto=Outro	D	0,0166	0,6522	10,4277
17 - UF=RS,Meio=0,Assunto=N040118	D	0,0054	0,6488	10,3723
18 - Meio=0,Assunto=N040118	D	0,0084	0,6465	10,3363
19 - Tipo=JF,Assunto=N040118	D	0,0062	0,5214	8,3354
20 - UF=RS,Assunto=N040118	D	0,0055	0,4621	7,3877
21 - Compe=Pr,Tipo=JF,Classe=C10822	D	0,0198	0,4374	6,9925
22 - Compe=Pr,Tipo=JF, Classe=C10977,Meio=1	C	0,0052	0,8846	5,3196
23 - Compe=Pr,Classe=C10977,Meio=1	C	0,0052	0,8810	5,2983
24 - Tipo=JF,Classe=C10977,Meio=1	C	0,0054	0,8598	5,1707
25 - Classe=C10977,Meio=1	C	0,0054	0,8566	5,1515
26 - Compe=Pr,Tipo=JF,Classe=C10977	C	0,0052	0,7864	4,7293
27 - Compe=Pr,Classe=C10977	C	0,0054	0,7503	4,5120
28 - Tipo=JF,Classe=C10977	C	0,0054	0,7365	4,4289
29 - Classe=C10977	C	0,0057	0,7054	4,2423

Continua...

Tabela 19 – Regras para atingir o segundo objetivo... continuação

Lado Esquerdo	Lado	Suporte	Confiança	Alavancagem
30 - Tipo=JE,UF=PR,Meio=0, Assunto=N04010202	C	0,0052	0,5563	3,3455
31 - Compe=Pr,Tipo=JE,Classe=C10828, UF=PR,Meio=0	C	0,0078	0,5502	3,3091
32 - UF=PR,Meio=0,Assunto=N04010202	C	0,0052	0,5470	3,2898
33 - Tipo=JE,Classe=C10828, UF=PR,Meio=0	C	0,0081	0,5336	3,2090
34 - Compe=Pr,Classe=C10828, UF=PR,Meio=0	C	0,0078	0,5124	3,0815
35 - Compe=Pr,Tipo=JE,UF=PR,Meio=0	C	0,0122	0,4707	2,8305
36 - Tipo=JF,Classe=C10828,UF=SC, Meio=1,Assunto=N040119	C	0,0061	0,4652	2,7974
37 - Classe=C10828,UF=SC,Meio=1, Assunto=N040119	C	0,0061	0,4619	2,7775
38 - Tipo=JF,Classe=C10828,UF=SC, Assunto=N040119	C	0,0061	0,4594	2,7628
39 - Tipo=JE,UF=PR,Meio=0	C	0,0128	0,4552	2,7374
40 - Classe=C10828,UF=SC, Assunto=N040119	C	0,0065	0,4420	2,6579
41 - Classe=C10828,UF=PR,Meio=0	C	0,0081	0,4343	2,6118
42 - Tipo=JF,Classe=C10992, Assunto=N040310	C	0,0093	0,4273	2,5696
43 - Classe=C10992,Meio=1, Assunto=N040310	C	0,0093	0,4262	2,5632
44 - Classe=C10992,Assunto=N040310	C	0,0096	0,4212	2,5328
45 - Tipo=JF,Classe=C10828, Meio=1,Assunto=N040119	C	0,0127	0,4143	2,4917
46 - Tipo=JF,Classe=C10992, Assunto=N040310	B	0,0098	0,4496	2,4658
47 - Tipo=JF,Classe=C10828, Assunto=N040119	C	0,0127	0,4093	2,4616
48 - Classe=C10992,Meio=1, Assunto=N040310	B	0,0098	0,4484	2,4596
49 - Classe=C10992,Assunto=N040310	B	0,0102	0,4476	2,4552
50 - Tipo=JF,UF=SC, Meio=1,Assunto=N040119	C	0,0077	0,4071	2,4482

Continua...

Tabela 19 – Regras para atingir o segundo objetivo... continuação

Lado Esquerdo	Lado	Suporte	Confiança	Alavancagem
51 - Compe=Pr,Tipo=JF,Classe=C10992, UF=SC,Meio=1	B	0,0099	0,4463	2,4479
52 - Classe=C10992,UF=RS, Assunto=N040104	B	0,0054	0,4416	2,4222
53 - Tipo=JF,UF=SC,Assunto=N040119	C	0,0077	0,4023	2,4192
54 - Compe=Pr,Tipo=JF, Classe=C10992,UF=SC	B	0,0099	0,4381	2,4029
55 - Tipo=JF,Classe=C10992, Meio=1,Assunto=N040104	B	0,0074	0,4221	2,3154
56 - Tipo=JF,Classe=C10992, Assunto=N040104	B	0,0074	0,4210	2,3092
57 - Tipo=JF,UF=SC,Assunto=N040310	B	0,0113	0,4174	2,2892
58 - Classe=C10992,Meio=1, Assunto=N040104	B	0,0075	0,4171	2,2880
59 - UF=SC,Meio=1,Assunto=N040310	B	0,0113	0,4168	2,2859
60 - UF=SC,Assunto=N040310	B	0,0115	0,4157	2,2800
61 - Classe=C10992,Assunto=N040104	B	0,0082	0,4144	2,2728
62 - Compe=Pr,Tipo=JF,Classe=C10992, UF=RS,Meio=1	B	0,0166	0,4092	2,2445
63 - Compe=Pr,Classe=C10992, UF=SC,Meio=1	B	0,0099	0,4079	2,2376
64 - Compe=Pr,Tipo=JF, Classe=C10992,Meio=1	B	0,0358	0,4037	2,2141
65 - Compe=Pr,Tipo=JF, Classe=C10992,UF=RS	B	0,0166	0,4009	2,1990
66 - Tipo=JF,Classe=C10992,Meio=1, Assunto=N040119	B	0,0073	0,4002	2,1949

Fonte: Dados da pesquisa

Estas regras são mais interessantes que as primeiras porque mostram, principalmente, os processos mais demorados, representados pela faixa de tempo D - mais de 2.180 dias. Estas regras apresentam suporte baixo, inferior a 4% e confiança de pelo menos 40%. Destas regras, a com maior suporte foi a de número 6 - processos oriundos da Justiça Federal, de meio físico, com duração superior a 2.180 dias. Dos processos físicos oriundos da Justiça Federal, 94,7% apresentaram tempo de atravessamento superior a 2.180 dias. Como observações podem ser destacadas:

- Destas regras com processos de tempo mais elevado, somente as regras número 14 e 15 possuem a competência administrativa do lado esquerdo e nenhuma regra apresentou, do lado esquerdo, a competência tributária. Por outro lado, as regras 2, 3, 8, 9, 21, 22, 23, 26, 27, 31, 34, 35, 51, 54, 62, 63, 64 e 65 referem-se à competência previdenciária, indicando que esta competência de processo é mais demorada. Por exemplo, a regra 3: processos físicos previdenciários oriundos da Justiça Federal do Rio Grande do Sul com tempo superior a 2180 dias representaram 2,62% do conjunto analisado e, dos processos físicos oriundos da Justiça Federal do Rio Grande do Sul, 97,6% apresentaram tempo de atravessamento superior a 2.180 dias;
- Quanto ao tipo de justiça de origem, apenas as regras número 33 e 35 contém a Justiça Estadual do lado esquerdo, e nenhuma das regras apresentou como tipo o TRF ou outro órgão;
- A regra número 29, de processos da classe C10977 (embargos infringentes), com faixa de tempo entre 1090 e 2180 dias representaram 0,6% do conjunto de dados, segundo a medida de suporte da regra. Mas dos processos da referida classe, 70,5% estavam dentro desta faixa de tempo de tempo de atravessamento.

Com estas regras, reforça-se a importância dos *outliers* no conjunto de dados: são estes casos menos frequentes que caracterizam os processos com perfil mais moroso. Como exemplo, podem ser citados os processos físicos, da classe agravo de instrumento e do assunto Aposentadoria por Tempo de Serviço: destes processos, 92,1% apresentaram tempo de atravessamento superior a 2180 dias, conforme a regra 7 da Tabela 19 .

As regras foram úteis para identificar o perfil dos processos mais demorados e as diferenças entre as covariáveis, corroborando com as conclusões da análise de sobrevivência e da análise descritiva. A seguir, apresenta-se a seção de discussão.

5.8 Discussão

Após as análises, os seguintes tópicos são discutidos: as variáveis, tanto a resposta como as covariáveis, o indicador usado, a aplicação de análise de sobrevivência e dos algoritmos, a qualidade do ajuste e a discussão sobre a realização de estudo prospectivo ou retrospectivo.

5.8.1 Variável resposta

Em relação à variável resposta, o tempo entre autuação e baixa dos processos, ou tempo de atravessamento, foi coletada e medida de forma quantitativa discreta. O tempo poderia ser medido através de outras escalas: quantitativa contínua, em que o tempo assume valores de dias, horas, minutos ou segundos. Houve também a tentativa de transformar a variável em uma escala ordinal (até 545 dias, de 545 a 1090 dias, de 1090 a 2180 dias e mais de 2180 dias).

A categorização permitiu, em vez de um valor numérico, uma faixa de tempo, um intervalo que pode ser de melhor visualização ou aplicação para quem recebe a informação. Por exemplo, a variável idade, quando tomada em faixas, pode ser mais esclarecedora que a idade propriamente dita. De forma análoga, a faixa de tempo de atravessamento pode dizer mais que um tempo médio ou mediano em dias.

Ou seja, a transformação facilitou a informação da duração processual e a tornou de fácil compreensão, incorporando a variabilidade encontrada na variável. Por outro lado, quando se analisa a variável de forma quantitativa discreta ou contínua, apoia-se em técnicas de análise como regressão, análise de variância, de sobrevivência, etc., melhor aplicadas a variáveis quantitativas.

5.8.2 Aplicabilidade dos algoritmos/técnicas para ajuste dos modelos

São feitas algumas considerações sobre as técnicas aplicadas referentes aos pressupostos, aos algoritmos e bibliotecas utilizados. Em relação à análise de sobrevivência, esta técnica pode ser considerada como a usual em estudos de tempo: esta ferramenta pode ser definida como o estudo de tempos de sobrevivência e os fatores que o influenciam, através de estudos observacionais prospectivos ou retrospectivos. Os objetivos da análise são estimar a função de sobrevivência, comparar os tempos de distribuição de tratamentos ou a elucidação dos fatores que influenciam os tempos, além da possibilidade de estimar modelos casuais ou preditivos.

Mesmo que outras técnicas possam ser usadas, o uso de análise de sobrevivência em estudos de tempo torna a pesquisa mais rica e compreensível à medida que possui uma gama de gráficos e testes para análise, além de estar disponível em diversos softwares. Mesmo que outras técnicas como regressão e análise de variância possam ser usadas, a análise de sobrevivência foi muito importante para a compreensão das covariáveis e sua influência no tempo mediano.

Por outro lado, há uma grande quantidade de dados (processos baixados) disponíveis para análise, cujas covariáveis já estão estruturadas. Esta quantidade de dados requer, além de um poder computacional maior, o desenvolvimento de métodos estatísticos que possibilite esta análise.

Assim, máquina de vetor suporte e redes neurais, por serem também direcionados a uma massa maior de dados, constituíram uma alternativa aos métodos estatísticos tradicionais. Tanto os algoritmos de máquina de vetor suporte e redes neurais são supervisionados e podem ser usados tanto para classificação como para a regressão, tornando-se assim, formas de proposição de modelos preditivos.

5.8.3 Previsibilidade de modelos e outras considerações

Sobre a capacidade de acerto dos modelos, é importante considerar alguns pontos, entre eles, a qualidade dos dados, a omissão de variáveis e a complexidade do fenômeno estudado. Pode ter havido problemas de qualidade de dados, incluindo valores ausentes, nulos, truncados,

incorretamente codificados ou corrompidos, como o caso do cadastramento do assunto dos processos, causando ruído aos modelos.

Em relação às covariáveis, a não inclusão daquelas consideradas importantes em outros trabalhos, como o valor da causa, pode influenciar o desempenho do modelo. Ressalta-se, também, o fato de haver apenas covariáveis categóricas, fazendo com que a variável resposta apresente previsões em níveis e não em valor contínuo.

O tempo de atravessamento final de um processo depende, além das variáveis mencionadas, de outros fatores não previsíveis no início do processo, como decisão no STF ou STJ, as alterações administrativas que implicam em mudança de órgão julgador, entre outras.

Além disso, o fenômeno pode ser considerado complexo, à medida que há uma diversidade de processos não criminais que englobam assuntos ambientais, de contratos, licitações, financiamento habitacional, etc. Uma alternativa à realização de um estudo geral, são pesquisas de temas específicos, como por exemplo, tempo de processos que tratam de aposentadorias ou de licitação.

Outro fator é a heterogeneidade da produção: uma das conclusões das entrevistas com especialistas é que cada Gabinete tem regras próprias para o sequenciamento da produção, ficando a cargo da gestão do gabinete o controle e o planejamento. Desta forma, é esperado que o tempo de atravessamento apresente alta variância, mesmo entre gabinetes de mesma competência e mesmo tipo de processo.

Outra ponderação é em relação às revisões do modelo. Alterações administrativas, como a implantação das turmas suplementares, alteram não apenas as medidas, como os modelos de predição. Assim, revisões periódicas são necessárias para atualizações ao longo do tempo.

5.8.4 Estudo retrospectivo ou prospectivo

A coleta de dados foi baseada naqueles processos que deixaram o estado de interesse, isto é, que foram baixados, sendo realizado, assim, um estudo retrospectivo. Os motivos da realização deste tipo de estudo, em vez de prospectivo, foi porque havia uma disponibilidade de informação sobre variáveis passadas, de fácil coleta, estruturadas e controlável em relação às datas de término e início do processo. Assim, o estudo foi conduzido do resultado para a causa, ou seja, do tempo para as covariáveis que o influenciaram.

Embora tenha sido realizado um estudo de coorte retrospectivo, pode-se discutir a realização de um prospectivo. Isto é, em vez de analisar processos findos, poderiam ter sido considerados processos iniciados em um determinado período e acompanhado estes processos até o término do estudo. Desta forma, os processos ainda não baixados no final do estudo seriam considerados censuras.

O objetivo seria incorporar as mudanças contemporâneas, tais como a implantação de turmas suplementares no modelo. Neste sentido, a primeira questão a ser discutida é a duração

do estudo. Para apresentar possíveis durações de um estudo prospectivo, a Tabela 20 mostra o percentual de processos civis no TRF4 conforme o ano de entrada e o ano de baixa.

Tabela 20 – Percentual de processos baixados segundo o ano de autuação do processo

Ano de baixa	Ano de autuação					
	2012	2013	2014	2015	2016	2017
2012	31,9	-	-	-	-	-
2013	25,3	34,2	-	-	-	-
2014	11,7	28,1	34,7	-	-	-
2015	7,9	10,1	26,3	34,4	-	-
2016	4,4	5,3	8,2	25,8	29,6	-
2017	5,0	7,1	10,0	14,3	32,8	28,6

Fonte: Dados fornecidos pela TRF4

Conforme a Tabela 20, para apresentar, no mínimo, 75% dos dados sem censura, o estudo deveria ser realizado por pelo menos quatro anos. Por exemplo, dos processos iniciados em 2014, 79,2% tinham sido baixados até dezembro de 2017, com tempo máximo de 1463 dias, porém, haveria 20,8% de processos com censura. Além disso, embora a realização de um estudo prospectivo incorpore as mudanças, corre o risco de não ser finalizado conforme o planejado.

Entretanto, sugere-se a construção de uma tábua de mortalidade de processos, tal como as utilizadas em estudos atuarias. No caso judicial, poderiam ser construídas tábuas de baixa de processos por competência e/ou meio processual. A sugestão é que as tábuas de baixa contenham as seguintes colunas:

x: coluna de idades ou tempo de atravessamento, em ordem cronológica e mensurado em meses;

l: quantidade de processos não baixados em cada idade;

d: quantidade de processos baixados em cada idade;

q: probabilidade de baixa em cada idade;

p: probabilidade de não baixar em cada idade;

e: expectativa completa de tempo de baixa para cada idade.

Caso haja interesse na realização de um estudo prospectivo e acredite-se que o risco de não finalização seja pequeno, pode-se comparar a qualidade da predição através da modelagem utilizando análise de sobrevivência e redes neurais. Sob este ponto de vista, haveria duas variáveis respostas: o tempo e censura, tanto na modelagem com redes neurais como em análise de sobrevivência.

A título de ilustração, sem o objetivo de comparar o desempenho dos modelos, foi realizada a comparação da análise do tempo, por meio de um estudo prospectivo, empregando análise de sobrevivência e redes neurais. Segue a descrição da comparação realizada:

- Variáveis respostas: i) tempo em dias entre a distribuição e a baixa de processos distribuídos em 2014 no TRF4. Para os processos não baixados até 31/12/2017, foi considerado o tempo entre a distribuição e esta data; ii) Censura, assumindo dois valores, respectivamente, 0 para processos não baixados até 31/12/2017 e 1, para baixados. No caso de processos baixados mais de uma vez, foi considerado o tempo na primeira baixa.

- Covariáveis: como o objetivo desta análise é mostrar a possibilidade de ajuste a partir destas duas técnicas, ou seja, ilustrativo, a análise foi restrita a covariável competência, com as categorias administrativa, previdenciária e tributária.

- Ajuste de um modelo de análise de sobrevivência: ajustado um modelo de regressão de tempo de vida acelerado, com o uso da distribuição lognormal, considerando censuras para processos não baixados;

- Ajuste de um modelo de regressão por redes neurais: o modelo ajustado possuiu duas variáveis respostas, respectivamente, o tempo de atravessamento transformado para o intervalo 0 a 1 e a censura. Foram usados os mesmos parâmetros do modelo anterior de redes neurais (2 camadas ocultas com 89 e 26 neurônios; taxa de aprendizado de 0,9; 288 interações e o algoritmo de retropropagação);

- O RMSE calculado para o modelo ajustado por análise de sobrevivência resultou em 510,8623, e o de redes neurais, em 760,0231. O objetivo desta comparação é mostrar que é possível ajustar um modelo a partir de redes neurais. Caso fosse realmente analisado, outras variáveis poderiam compor o modelo, além da busca de parâmetros melhores, o que, provavelmente, resultaria em diferentes valores para o RMSE, provavelmente, menor.

5.9 Estabelecimento de padrões de tempo de atravessamento

Em relação ao estabelecimento de padrões aceitáveis para o tempo de duração processual, comparam-se como seriam os prazos razoáveis caso fossem definidos segundo o indicador *disposition time*, conforme usado pelo CEPEJ. O indicador *disposition time*, DT, é dado por:

$$DT = 365 \frac{\text{Casos pendentes no final do ano}}{\text{Casos resolvidos no ano}} \tag{72}$$

Considerando os casos pendentes no final do período, do indicador *disposition time*, como o WIP; e os casos encerrados no período como a taxa de saída (*throughput*), multiplicado por 365, o tempo obtido por meio do indicador *disposition time* é igual ao CT.

Desta forma conclui-se que o *disposition time*, conforme proposto pelo CEPEJ, foi inspirado na Lei de Little e que pode ser estabelecido pela gerência.

Como demonstração da aplicação de lei de Little para o estabelecimento de padrões, calcula-se o indicador *disposition time* para o caso do TRF4, como proposta de parâmetros de tempo razoável. Porém, como no caso do CEPEJ, o indicador é de acordo com a classe e a competência, não de forma geral.

Para o caso do TRF4, é sugerido um indicador adaptado, substituindo os casos pendentes pela quantidade de processos em tramitação e os casos resolvidos, por processos baixados. Contudo, para a adaptação, são analisadas, em um primeiro momento, as premissas para a aplicação da lei.

Quanto à primeira premissa, de conservação do fluxo, esta pode ser observada pelas taxas de saída e de entrada, podendo ser observada por meio da Figura 3, que fornece a entrada (distribuição) e saída (baixa) de processos. Como a entrada é superior à saída na maioria dos períodos, esta desigualdade pode influenciar a premissa de que o trabalho em processo (WIP) tenha aproximadamente o mesmo tamanho no início e no final do intervalo de tempo. Assim, as premissas necessitam ser avaliadas periodicamente.

A segunda premissa, de que não há itens perdidos no sistema, pode ser comprovada pelo fato de que todo o processo que entra no sistema deve ser baixado, mesmo que tenha sua distribuição cancelada. Isto é, todo processo que entra, sai, com o devido tratamento.

Para mostrar estes tempos razoáveis no TRF4, a Tabela 21 compara o indicador às medidas de tempo de atravessamento das duas classes mais frequentes, considerando apenas os processos eletrônicos, baixados no ano de 2017.

Tabela 21 – Tempo de atravessamento observado e *disposition time* das classes apelação civil e agravo de instrumento

Classe e competência	*Disposition time*	Tempo de atravessamento		
		Mediano	Médio	3º quartil
Apelação civil				
Administrativa	494	352	618,7	882
Tributária	571	161	470,5	611
Previdencária	502	580	732,5	1143
Agravo de Instrumento				
Administrativa	288	151	242,5	234
Tributária	208	142	227,2	216
Previdencária	194	127	196,3	177

Fonte: Dados da pesquisa

A exceção de processos previdenciários da classe de apelação cível, o indicador *disposition time* é maior que o valor mediano para todas as competências e classes. Excetuando, dos agravos de instrumento, os processos administrativos e previdenciários, o valor do indicador é menor que o terceiro quartil. Assim, caso o *disposition time* fosse utilizado para estabelecer padrões de tempo razoável, o indicador *case backlog*, ou o percentual de processos resolvidos fora do prazo, estaria entre 25% e 50% para estes casos. Isto é, em pelo menos 25% dos processos, não seria alcançado o tempo razoável. Este índice de *case backlog* é obtido porque a variabilidade é alta, conforme pode ser observado no apêndice C.

Desta forma, alerta-se para a possibilidade de não cumprimento do indicador *case backlog*, se o *disposition time* for usado para estabelecimento de tempo razoável de processo.

A partir do estabelecimento de padrões, ações que visam à redução da variabilidade do tempo de processos podem ser elaboradas, com o intuito de diminuir o indicador *case backlog*: à medida que o desvio padrão do tempo é reduzido, a previsão se torna mais próxima do valor real, diminuindo o número de processos com tempo superior ao estipulado, reduzindo consequentemente, o indicador. Estas ações incluem revisão do fluxo dos processos de trabalho, com alterações nos mapas processuais, por exemplo.

O estabelecimento de padrões de tempo de atravessamento, para o Poder Judiciário, é importante porque:

- contribui para a transparência, uma vez que estabelece prazos a serem cumpridos e permite a mensuração do não cumprimento;
- fornece informações para o planejamento de recursos, principalmente, humanos. Também fornece informações para as partes processuais, auxiliando no estabelecimento de faixas de tempo;
- o uso de modelos de previsão permite o cálculo da probabilidade do possível não cumprimento do prazo proposto. Por exemplo, se o prazo fixado for igual a 365 dias, e o tempo previsto para aquele processo é superior a este prazo, pode-se marcar" o processo, como um *poka-yoke*, a fim de acompanhá-lo para um possível não cumprimento. Para tanto, um modelo com a análise a partir da variável resposta quantitativa deve ser usado.

Neste sentido, o modelo de previsão fornece uma informação a ser usada tanto para a gestão judiciária como para as partes. Informando as partes sobre os tempos processuais, estas podem decidir em conciliar em vez de litigar. Desta forma, há redução na taxa de entrada com consequente redução dos processos pendentes. Com menos processos pendentes, acredita-se que o tempo seja menor.

Além disso, a informação sobre o tempo é importante para o estabelecimento de metas e para o planejamento de recursos, como humanos, de equipamentos, etc.

A percepção do tempo menor, por parte dos jurisdicionados, pode influenciar em outros indicadores, tais como a satisfação com os serviços prestados. Como sugestão de novos indicadores, além do *disposition time*, sugere-se a adoção do indicador *case backlog*, e os indicadores europeus derivados relacionados, como o *case turnover ratio*, ou a razão de casos resolvidos em relação aos casos não resolvidos, o qual mede o quanto falta para baixar o estoque.

Em resumo, uma contribuição extra desta tese, não mencionada no Capítulo 1, está na aplicação da lei de Little para o estabelecimento de padrões razoáveis para o tempo de atravessamento processual. De qualquer forma, o uso da Lei de Little para o estabelecimento de parâmetros deve ser revisto periodicamente, principalmente, no tocante às premissas de igualdade de taxas de entrada e de saída e a quantidade de casos pendentes iguais ao início e final do período.

6 CONSIDERAÇÕES FINAIS

No Brasil, há uma preocupação incipiente com a eficiência dos serviços públicos por parte de órgãos como o Ministério do Planejamento, Desenvolvimento e Gestão e o Conselho Nacional de Justiça, evidenciadas pela criação do Programa Nacional de Gestão Pública e Desburocratização - Gespública, da Carta de Serviços e das metas do Poder Judiciário.

O Ministério da Justiça reconhece os problemas da justiça brasileira, apontando o alto número de processos em estoque, a falta de acesso à justiça e a morosidade como os principais problemas do Poder Judiciário. A morosidade, entretanto, é considerada como o principal aspecto da chamada crise do judiciário.

Neste sentido, esta tese propôs estruturar e comparar modelos de previsão de tempo de atravessamento processual de ações judiciais civis na justiça federal de 2^a Instância, para servir de informação para as partes processuais e administração. O estudo foi realizado no Tribunal Regional Federal da 4^a Região, órgão que apresenta os problemas de morosidade e do alto número de processos.

Foram utilizados os dados do *datawarehouse* do TRF4 sobre processos cíveis baixados em 2017. No banco de dados institucional, há variáveis disponíveis referentes a classe, assunto, justiça de origem, unidade da federação de origem, meio processual. A quantidade de processos baixados no período de um ano têm sido superior a 100.000, a qual exige um poder computacional adicional, que por sua vez estimula o desenvolvimento de métodos como mineração de dados e aprendizado de máquina para análise.

Para alcançar o objetivo geral proposto, foram ajustados quatro modelos para o tempo de atravessamento. O primeiro modelo foi ajustado através de redes neurais para regressão com o uso do algoritmo retropropagação. O segundo utilizou máquina de vetor suporte para regressão, através da biblioteca Libsvm. O modelo utilizado para comparação com os dois citados, o terceiro modelo, foi ajustado mediante aplicação de análise de sobrevivência, considerada como técnica clássica e habitual para análise em estudos quantitativos envolvendo tempos.

A variável resposta usada foi o tempo em dias entre a autuação do processo no 2^o Grau e a baixa. Esta variável foi transformada em um valor entre 0 e 1 para o ajuste do modelo de redes neurais, e para máquina de vetor suporte, a variável foi normalizada.

Para comparar o desempenho dos modelos ajustados por máquina de vetor suporte e redes neurais, foi calculado o RMSE - Raiz quadrada média do erro de predição - a partir dos valores preditos do conjunto de teste.

O RMSE dos dois modelos pode ser considerado alto em comparação ao tempo médio, de 790 dias: respectivamente, 470,2745 e 466,3132 para o modelo ajustado por redes neurais e

máquina de vetor suporte.

Os modelos de máquina de vetor suporte e redes neurais apresentaram valores de RMSE muito próximos, embora o modelo ajustado com o uso de máquina de vetor suporte tenha apresentado resultado superior, isto é, RMSE menor. Porém, o modelo ajustado por análise de sobrevivência não demonstrou desempenho superior aos demais, embora tenha sido útil para a análise, para a compreensão das variáveis e a influência destas no tempo mediano.

De qualquer forma, de acordo com a literatura, em aprendizado de máquina é extremamente raro um modelo generalizar perfeitamente a cada caso imprevisto, em parte devido ao ruído, causados por eventos aparentemente aleatórios, como erros de medida e o fenômeno em estudo ser um assunto complexo.

Além disso, considerando que um modelo é uma simplificação, uma representação conceitual do mundo real, é provável que informações relevantes não tenham sido incluídas. Para diminuir o valor do RMSE, outras variáveis como valor da causa, poderiam ser incluídas ou ainda a realização da análise por tema, como exemplo, aposentadorias.

Um quarto modelo foi então ajustado com a variável resposta discretizada, isto é, transformada em ordinal por meio da estratificação em faixas de tempo. Para este ajuste, foi utilizada máquina de vetor suporte para classificação, através da biblioteca Libsvm. Este modelo apresentou uma acurácia total de 69,7%.

Após os ajustes, foram aplicadas regras de associação às faixas de tempo com o objetivo de encontrar as características dos processos mais lentos e morosos. As conclusões obtidas com estas regras corroboraram o que tinha sido obtido com a análise de sobrevivência e a análise descritiva: os processos previdenciários são mais lentos em comparação às demais competências, os agravos de instrumento são mais rápidos que apelações cíveis, assim como os processos eletrônicos são mais rápidos que os físicos.

Todos os modelos foram obtidos com a utilização do software R, com os pacotes de *neuralnet* e *deepnet* para redes neurais, *e1071* para máquinas de vetor suporte, *survival* e *flexsurv* para análise de sobrevivência. Para as regras de associação, foi usado o algoritmo *apriori* através do pacote *arules*.

A partir do ajuste destes modelos, algumas discussões foram realizadas. Acredita-se que as técnicas utilizadas aplicam-se aos dados, destacando a aplicação de análise de sobrevivência para a obtenção de resultados referentes à influência das variáveis e máquina de vetor suporte (classificação) para obter estimativas da faixa de tempo. Embora a discretização reduza as técnicas de análise, a informação sobre uma faixa de tempo pode ser mais esclarecedora que uma medida de tempo exato.

Quanto ao estabelecimento de padrões para consideração de prazos excessivos, o indicador *disposition time*, como utilizado pelo CEPEJ, não teria bons resultados no caso estudado. Para as duas classes de processos eletrônicos analisados, apelação cível e agravo de instru-

mento, o indicador de *case backlog*, ou a porcentagem de casos com tempo maior do que prazos estabelecidos, estaria entre 25 e 50%.

Assim, duas sugestões podem ser feitas quanto à divulgação do tempo de atravessamento: divulgar a faixa de tempo provável, principalmente, para informação às Partes processuais e, para o estabelecimento de parâmetros, usar a previsão do tempo obtido ou pelo modelo de redes neurais ou por máquina de vetor suporte, pois ambos tiveram desempenho semelhante.

Sugere-se a realização dos seguintes trabalhos referentes à medição de tempo de atravessamento processual:

- Realização de um trabalho prospectivo com a construção de tábuas de vida;
- A agregação de outras variáveis no modelo como valor da causa e prioridades de julgamento;
- A modelagem de questões individulizadas, como a aposentadoria, contratos específicos, etc;
- O estabelecimento de padrões com a utilização de medidas diferentes da média, como o terceiro quartil;
- A utilização de outras técnicas tais como regressão quantílica e outros algoritmos para a classificação como florestas aleatórias.

A partir do estabelecimento de padrões, ações que visam à redução da variabilidade do tempo de processos podem ser elaboradas, com o intuito de diminuir o tempo de atravessamento. Estas ações incluem revisão do fluxo dos processos de trabalho, com alterações nos mapas processuais, por exemplo. Além disso, a informação sobre o tempo é importante para o estabelecimento de metas e para o planejamento de recursos, como humanos, de equipamentos, etc.

Por fim, as contribuições propostas atingidas:

- Levantamento de trabalhos realizados sobre tempo de atravessamento processual, o que mostrou os trabalhos realizados sobre o tema, as técnicas e variáveis usadas.
- Análise do tempo de atravessamento por meio de variável resposta discretizada, análise do tempo na justiça federal e ajuste por meio de máquina de vetor suporte para regressão, o que não havia sido realizado até então no Brasil;
- Análise das variáveis que compõem a lei de Little e a viabilidade de estabelecimento de parâmetros de tempo para o caso brasileiro, baseado nestas variáveis. Esta foi uma contribuição extra, não planejada no primeiro momento;

- Contribuir para os estudos de jurimetria, um campo emergente sobre métodos quantitativos aplicado ao direito.

Em resumo, o assunto abordado é de fundamental importância para o país à medida que a morosidade é considerada um empecilho ao crescimento econômico e a má atuação do Poder Judiciário tem reflexos na economia do país.

Além disso, há um cenário propício para o uso de indicadores e de modelos de predição: o Ministério do Planejamento e Gestão apoia a tomada de decisões e a execução de ações suportadas por medição e análise do desempenho, levando-se em consideração as informações disponíveis.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, R. S. de. *Anotações aos Artigos 284 a 290. In: Novo Código de Processo Civil Anotado - OAB/RS*. 2016. 227–231 p. Disponível em: <http://www.oabrs.org.br/novocpcanotado/novo_cpc_anotado_2015.pdf>. Acesso em: 12 abril 2016. Citado na página 33.

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: ACM. *Acm sigmod record*. [S.l.], 1993. v. 22, n. 2, p. 207–216. Citado na página 93.

ALCANTARA, F. M. M. d. et al. Análise informétrica da constituição federal brasileira. *Revista Gestão e Conhecimento on line*, Facet Faculdade, v. 8, n. 1, p. 117–127, 2014. Citado 2 vezes nas páginas 49 e 52.

ALLISON, P. D. *Survival analysis using SAS: a practical guide*. Second. North Carolina: Sas Institute, 2010. 327 p. Citado 3 vezes nas páginas 26, 85 e 99.

BALDAN, G. R. *Meio eletrônico: uma das formas de diminuição do tempo de duração do processo no 4º Juizado Especial Cível de Porto Velho/RO*. 172 p. Dissertação (Dissertação de Mestrado. (Mestrado Profissional em Poder Judiciário)) — Fundação Getúlio Vargas, Rio de Janeiro, 2011. Citado 2 vezes nas páginas 20 e 33.

BIELEN, S.; MARNEFFE, W.; VEREECK, L. An empirical analysis of case disposition time in belgium. *Review of Law & Economics*, De Gruyter, v. 11, n. 2, p. 293–316, 2015. Citado na página 21.

BONOTTO, C. L. F. *Política nacional de conciliação: política pública implementação pelo Conselho Nacional de Justiça*. 119 p. Dissertação (Dissertação de Mestrado. (Mestrado Profissional em Poder Judiciário)) — Fundação Getúlio Vargas, Rio de Janeiro, 2012. Citado na página 37.

BORDASCH, R. W. d. S. *Gestão cartorária: controle e melhoria para a razoável duração dos processos*. 138 p. Dissertação (Dissertação de Mestrado. (Mestrado Profissional em Poder Judiciário)) — Fundação Getúlio Vargas, Rio de Janeiro, 2009. Citado na página 20.

BOX, G. E.; JENKINS, G. M. *Time series analysis: forecasting and control, revised ed*. [S.l.]: Holden-Day, 1976. Citado na página 30.

BRASIL. *Constituição (1988). Constituição da República Federativa do Brasil de 1988.* 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm>. Acesso em: 18 mai. 2016. Citado 2 vezes nas páginas 14 e 32.

BRASIL. *LEI Nº 13.105, DE 16 DE MARÇO DE 2015. Código de Processo Civil.* 2015. Disponível em: $< http : //www.planalto.gov.br/ccivil_03/_{ato}2015 - 2018/2015/lei/l13105.htm >$. Acesso em 18 mai. 2016. Citado na página 34.

BRASIL. *Emenda Constitucional nº45, de 30 de dezembro de 2004. Altera dispositivos dos arts. 5º, 36, 52, 92, 93, 95, 98, 99, 102, 103, 104, 105, 107, 109, 111, 112, 114,*

115, 125, 126, 127, 128, 129, 134 e 168 da Constituição Federal, e acrescenta os arts. 103-A, 103B, 111-A e 130-A, e dá outras providências. 2016. Dispoível em: $< http : //www.planalto.gov.br/ccivil_03/constituicao/emendas/emc/emc45.htm >$. Acesso em 18 abr. 2016. Citado 2 vezes nas páginas 27 e 33.

BRASIL. *LEI No 10.259, DE 12 DE JULHO DE 2001. Dispõe sobre a instituição dos Juizados Especiais Cíveis e Criminais no âmbito da Justiça Federal.* 2016. Disponível em: $< http : //www.planalto.gov.br/ccivil\_03/leis/LEIS\_2001/L10259.htm >$. Acesso em 18 mai. 2016. Citado na página 36.

BRASIL, G. O abc da matemática atuarial e princípios gerais de seguros. *Porto Alegre: Sulina*, 1985. Citado na página 92.

BRESSER-PEREIRA, L. C. Burocracia pública na construção do brasil. *São Paulo*, 2008. Citado na página 27.

BUSCAGLIA, E.; ULEN, T. A quantitative assessment of the efficiency of the judicial sector in latin america. *International Review of Law and Economics*, Elsevier, v. 17, n. 2, p. 275–291, 1997. Citado na página 38.

CARVALHO, J. de; GONÇALVES, W. F.; OLIVEIRA, S. A. de. *A morosidade da justiça brasileira nos julgamentos dos processos de pessoas idosas*. 2012. 1257–1294 p. Citado 2 vezes nas páginas 36 e 37.

CAUCHICK, P. A. C.; FLEURY, A. C. C. *Metodologia de pesquisa em engenharia de produção e gestão de operações*. Second. Rio de Janeiro: Elsevier, 2012. 260 p. Citado na página 98.

CAVALCANTE, A. R. *A morosidade no poder judiciário brasileiro e suas causas possíveis*. 49 p. Dissertação (Monografia. (Especialização em Administração Judiciária)) — Universidade Estadual Vale do Acaraúva, Fortaleza, 2008. Citado na página 36.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, Acm, v. 2, n. 3, p. 27, 2011. Citado na página 117.

CHAPHALKAR, N.; IYER, K.; PATIL, S. K. Prediction of outcome of construction dispute claims using multilayer perceptron neural network model. *International Journal of Project Management*, Elsevier, v. 33, n. 8, p. 1827 – 1835, 2015. ISSN 0263-7863. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0263786315001416>. Citado na página 49.

CHAPPE, N. Demand for civil trials and court congestion. *European Journal of Law and Economics*, Springer, v. 33, n. 2, p. 343–357, 2012. Citado 2 vezes nas páginas 36 e 37.

CLEMES, J. G. *Método para funcionamento eficiente e eficaz de uma unidade judiciária: como a representação dos processos de trabalho por meio de fluxogramas pode revolucionar a prestação do serviço jurisdicional nos Juizados Especiais Cíveis*. 92 p. Dissertação (Dissertação de Mestrado. (Mestrado Profissional em Poder Judiciário)) — Fundação Getúlio Vargas, Rio de Janeiro, 2010. Citado na página 20.

CNJ. *Resolução n^o 76/2009, de 12 de maio de 2009.* 2009. Disponível em: $< http : //www.cnj.jus.br >$. Acesso em 02 mai. 2016. Citado 5 vezes nas páginas 19, 20, 43, 44 e 46.

CNJ. *Primeira instância, segunda instância... Quem é quem na Justiça brasileira?*. 2010. Disponível em: $< http : //cnj.jus.br/noticias/cnj/59220 - primeira - instancia - segunda - instancia - quem - e - quem - na - justica - brasileira >$. Brasília, 1^o out. 2010. Acesso em 09 jun. 2016. Citado 2 vezes nas páginas 32 e 33.

CNJ. *Pesquisa de Clima Organizacional e Satisfação dos Usuários*. Brasília, 2011. Disponível em <http://http://www.cnj.jus.br/gestao-e-planejamento/gestao-e-planejamento-do-judiciario/pesquisa-de-satisfacao-e-clima-organizacional>. Acesso em 30 jun. 2016. Citado na página 45.

CNJ. *Ministério da Justiça aponta três principais problemas do Judiciário*. 2014. Disponível em: <http://www.cnj.jus.br/noticias/cnj/61341-ministerio-da-justica-aponta-tres-principais-problemas-do-judiciario>. Acesso em 15 mar. 2016. Citado 3 vezes nas páginas 14, 23 e 101.

CNJ. *Quem somos, visitas e contatos*. Brasília: [s.n.], 2016. Disponível em <http://cnj.jus.br/sobre-o-cnj/quem-somos-visitas-e-contatos>. Acesso em 02 jun. 2016. Citado na página 33.

CNJ. Relatório justiça em números, elaborado pelo conselho nacional de justiça. *http://www. cnj. jus. br/files/conteudo/arquivo/2016/10/b8f46be3dbbff344931a933579915488. pdf. Consulta em*, v. 10, n. 11, p. 2016, 2016. Citado na página 24.

CNJ. *Justiça em Números 2017: ano base 2016*. Brasília, 2017. Citado na página 14.

COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência aplicada. In: *ABE-Projeto Fisher*. São Paulo: Edgard Blücher, 2006. p. 367. Citado 6 vezes nas páginas 84, 85, 89, 91, 92 e 99.

COUTO, M. B.; OLIVEIRA, S. P. de. Gestão da justiça e do conhecimento: a contribuição da jurimetria para a administração da justiça. *Revista Jurídica*, v. 2, n. 43, p. 771–801, 2016. Citado na página 21.

CUNHA, A. d. S. C. et al. *Custo unitário do processo de execução fiscal na Justiça Federal: relatório de pesquisa*. Brasília, 2011. Citado na página 36.

CUNHA, L. G. et al. *Relatório ICJBrasil- 2^o semestre/2015*. São Paulo, 2015. Citado na página 28.

DAKOLIAS, M. *O setor judiciário na América Latina e no Caribe: elementos para reforma*. New York, 1996. Citado na página 27.

DALTON, T.; SINGER, J. M. Bigger isn't always better: An analysis of court efficiency using hierarchical linear modeling. *Pace L. Rev.*, HeinOnline, v. 34, p. 1169, 2014. Citado 3 vezes nas páginas 19, 21 e 102.

ERICKSEN, P.; STOFLET, N.; SURI, R. Manufacturing critical-path time (mct): the QRM metric for lead time. *Manufacturing Critical-path Time (MCT): the QRM metric for lead time*, Center for QRM Wisconsinˆ eMadison, Wisconsin-Madison, 2007. Citado na página 47.

ESMAFE-PR. *Justiça Federal ao alcance de todos*. Curitiba, 2016. Citado na página 35.

FABRI, M.; CARBONI, N. *Saturn guidelines for Judicial Time Management: Comments and Implementation Examples, COE/CEPEJ*. 2013. Citado 2 vezes nas páginas 15 e 23.

FALAVIGNA, G. et al. Judicial productivity, delay and efficiency: A directional distance function (ddf) approach. *European Journal of Operational Research*, Elsevier, v. 240, n. 2, p. 592–601, 2015. Citado 2 vezes nas páginas 27 e 47.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From data mining to knowledge discovery in databases*. 1996. 37–54 p. Citado 3 vezes nas páginas 64, 65 e 70.

FEITOSA, A. A. C. *Do poder judiciário: a morosidade no âmbito da justiça estadual*. 58 p. Dissertação (Monografia. (Especialização em Administração Judiciária) — Universidade Estadual do Vale do Acaraú, Fortaleza, 2007. Citado na página 37.

FERREIRA, A. R. Modelo de excelência em gestão pública no governo brasileiro: importância e aplicação. In: *XIV Congreso Internacional del CLAD sobre la Reforma del Estado y de la Administración Pública, Salvador de Bahia, Brasil*. [S.l.: s.n.], 2009. p. 27–30. Citado na página 15.

FRITSCH, S.; GUENTHER, F.; GUENTHER, M. F. *Package 'neuralnet' Training of neural networks*. 2012. Citado 2 vezes nas páginas 115 e 117.

FRITSCH, S.; GUENTHER, F.; GUENTHER, M. F. Package 'neuralnet'. *The Comprehensive R Archive Network*, 2016. Citado na página 117.

GEORGE, T. E.; GUTHRIE, C. Induced litigation. *Nw. UL Rev.*, HeinOnline, v. 98, n. 2, p. 545–578, 2003. Citado 2 vezes nas páginas 37 e 38.

GESPúBLICA. Carta de serviços ao cidadão. *Guia Metodológico. Brasília: SEGEP/MP*, 2014. Citado 2 vezes nas páginas 16 e 17.

GESPÚBLICA, P. Documento de referência. cadernos gespública. *Brasília: SEGEP/MP*, 2009. Citado na página 15.

GESPÚBLICA, P. Modelo de excelência em gestão pública. *Brasília: SEGEP/MP*, 2014. Citado 4 vezes nas páginas 15, 16, 18 e 28.

GIRÃO, A. C. *Justiça a Passos Lentos. Aspectos sobre a morosidade da Justiça*. 62 p. Dissertação (Monografia. (Especialização em Administração Judiciária)) — Universidade Estadual Vale do Aracaú, Fortaleza, 2008. Citado 2 vezes nas páginas 36 e 37.

GRIBEL, D. L. *A clustering-based approach to detect probable outcomes of lawsuits*. 60 p. Dissertação (Tese. (Doutorado em Informática Aplicada)) — Universidade Federal do Estado do Rio de Janeiro - UNIRIO, 2014. Citado 4 vezes nas páginas 49, 51, 54 e 59.

GRUGINSKIE, L. A. dos S.; VACCARO, G. L. R. Lawsuit lead time prediction: Comparison of data mining techniques based on categorical response variable. *PloS one*, Public Library of Science, v. 13, n. 6, p. e0198122, 2018. Citado 7 vezes nas páginas 22, 29, 49, 53, 54, 60 e 160.

GUIMARÃES, J. L. V. Z. *Relação entre o princípio da eficiência administrativa e o da razoável duração do processo no âmbito da reforma do judiciário*. 58 p. Dissertação (Monografia. (Especialização em Administração Judiciária)) — Escola Superior da Magistratura do Estado do Ceará, Fortaleza, 2014. Citado na página 20.

GÜNTHER, F.; FRITSCH, S. neuralnet: Training of neural networks. *The R journal*, v. 2, n. 1, p. 30–38, 2010. Citado 3 vezes nas páginas 79, 81 e 84.

HAHSLER, M. et al. Package ‘arules’. 2018. Citado na página 119.

HALL, D.; KEILITZ, I. *Global Measures of Court Performance*. 2012. Citado 2 vezes nas páginas 42 e 46.

HALL, M.; WITTEN, I.; FRANK, E. *Data mining: Practical machine learning tools and techniques*. Third. Burlington: Kaufmann, 2011. 665 p. Citado 9 vezes nas páginas 65, 71, 72, 76, 78, 79, 82, 93 e 94.

HAMILTON, B. A. *The field guide to data science*. Web links: Booze Allen Hamilton Inc., 2013. 123 p. Citado na página 62.

HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. Second. San Francisco: Elsevier, 2006. 772 p. Citado 8 vezes nas páginas 64, 65, 66, 68, 69, 70, 71 e 121.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. Boston: Elsevier, 2011. Citado 12 vezes nas páginas 26, 64, 72, 73, 80, 81, 82, 83, 84, 112, 113 e 114.

HANSON, R.; OSTROM, B.; KLEIMAN, M. The pursuit of high performance. *International Journal for Court Administration*, International Association for Court Administration, v. 3, n. 1, p. 2–13, 2010. Citado 3 vezes nas páginas 40, 41 e 46.

HARDIN, G. The tragedy of the commons. *Science*, American Association for the Advancement of Science, v. 162, n. 3859, p. 1243–1248, 1968. Citado na página 37.

HASTIE, T. *Package ‘gam’*. 2016. Citado 2 vezes nas páginas 78 e 115.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Unsupervised learning. In: *The elements of statistical learning*. [S.l.]: Springer, 2009. p. 485–585. Citado 12 vezes nas páginas 26, 72, 73, 74, 75, 76, 77, 79, 80, 81, 82 e 83.

HOPP, W. J.; SPEARMAN, M. L. *A ciência da fábrica*. Third. Porto Alegre: Bookman, 2013. 720 p. Citado 2 vezes nas páginas 47 e 48.

HORNIK, K.; MEYER, D.; KARATZOGLOU, A. Support vector machines in r. *Journal of statistical software*, American Statistical Association, v. 15, n. 9, p. 1–28, 2006. Citado 2 vezes nas páginas 26 e 74.

HSU, C.-W. et al. A practical guide to support vector classification. Taipei, 2003. Citado na página 75.

JACKSON, C.; JACKSON, M. C. *Package ‘flexsurv’*. 2017. Citado 2 vezes nas páginas 115 e 116.

JENKINS, S. P. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 2005. Citado na página 92.

JOHNSON, N. L. *Survival models and data analysis*. Toronto: John Wiley & Sons, 1999. v. 74. 457 p. Citado na página 84.

JONSKI, K.; MANKOWSKI, D. Is sky the limit? revisiting ‘exogenous productivity of judges’ argument. *International Journal for Court Administration*, International Association for Court Administration, v. 6, n. 2, p. 53–72, 2014. Citado na página 37.

JR, D. W. H.; LEMESHOW, S. Applied survival analysis: regression modelling of time to event data (1999). *Eur Orthodontic Soc*, p. 561–2, 1999. Citado na página 92.

JR, E. S. *Modelagem de sistema de conhecimento para apoio a decisão sentencial na Justiça Estadual*. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2012. Citado 3 vezes nas páginas 33, 49 e 50.

JR, I. T. G. *A Tragédia do Judiciário: subinvestimento em capital jurídico e sobreutilização do Judiciário*. 163 p. Dissertação (Tese. (Doutorado em Economia)) — Universidade de Brasília-UnB, 2012. Citado na página 14.

JUNIOR, P.; ARRAIS, J. *Seleção de covariáveis para modelos de sobrevivência via verossimilhança penalizada*. Tese (Doutorado) — Universidade de São Paulo, 2009. Citado na página 92.

KIRAT, T. et al. Performance-based budgeting and management of judicial courts in france: an assessment. *International Journal for Court Administration*, International Association for Court Administration, v. 2, n. 2, p. 12–20, 2010. Citado 3 vezes nas páginas 41, 42 e 46.

KIRK, M. *Thoughtful Machine Learning: A Test-driven Approach*. First. Cambridge: O'Reilly Media, Inc., 2014. 235 p. Citado 2 vezes nas páginas 70 e 71.

KLEIN, J. P.; MOESCHBERGER, M. L. *Survival analysis: techniques for censored and truncated data*. [S.l.]: Springer Science & Business Media, 2005. Citado 7 vezes nas páginas 85, 87, 88, 89, 90, 91 e 92.

KOH, Y. S.; RAVANA, S. D. Unsupervised rare pattern mining: a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, ACM, v. 10, n. 4, p. 45, 2016. Citado na página 95.

KOH, Y. S.; ROUNTREE, N. Finding sporadic rules using apriori-inverse. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2005. p. 97–106. Citado na página 95.

LANTZ, B. *Machine learning with R*. [S.l.]: Packt Publishing Ltd, 2013. Citado 5 vezes nas páginas 79, 80, 82, 83 e 84.

LEIRIA, M. L. L. Limites no juízo de admissibilidade dos recursos especial e extraordinário. *Revista de Doutrina da 4ª Região*, Porto Alegre, jul 2006. Citado na página 32.

LEPORE, L.; METALLO, C.; AGRIFOGLIO, R. Evaluating court performance: Findings from two italian courts. *International Journal for Court Administration*, International Association for Court Administration, v. 4, n. 3, p. 82–93, 2012. Citado 3 vezes nas páginas 19, 40 e 46.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. *Mining of massive datasets*. [S.l.]: Cambridge university press, 2014. Citado na página 72.

LEWIS, N. Deep learning made easy with r. *with RA Gentle Introduction to Data Science*, p. 6, 2016. Citado 3 vezes nas páginas 66, 67 e 84.

LIMAS, M. C. et al. *Package 'AMORE'. A more flexible neural network package*. 2010. Disponível em: <http://rwiki.sciviews.org/doku.php?id=packages:cran:amore>. Citado na página 115.

LITTLE, J. D.; GRAVES, S. C. Little's law. In: *Building intuition*. [S.l.]: Springer, 2008. p. 81–100. Citado 2 vezes nas páginas 30 e 48.

LIU, B.; HSU, W.; MA, Y. Pruning and summarizing the discovered associations. In: ACM. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 1999. p. 125–134. Citado na página 96.

LIU, X. *Survival analysis: models and applications*. West Sussex: John Wiley & Sons, 2012. 446 p. Citado 3 vezes nas páginas 26, 84 e 85.

LIU, Y.-H.; CHEN, Y.-L.; HO, W.-L. Predicting associated statutes for legal problems. *Information Processing & Management*, Elsevier, v. 51, n. 1, p. 194–211, 2015. Citado na página 72.

LOEVINGER, L. Jurimetrics–the next step forward. *Minn. L. Rev.*, HeinOnline, v. 33, p. 455, 1948. Citado na página 21.

LUNA, J. M.; ROMERO, J. R.; VENTURA, S. On the adaptability of g3parm to the extraction of rare association rules. *Knowledge and information systems*, Springer, v. 38, n. 2, p. 391–418, 2014. Citado na página 95.

MARTINS, D. B. *A provisão de serviços públicos de resolução judicial de litígios: análise económica do sistema judicial português*. Tese (Doutorado) — Instituto Superior de Economia e Gestão, 2009. Citado na página 15.

MARTINS, H. F.; MARINI, C. Um guia de governança para resultados na administração pública. In: *Um guia de governança para resultados na administração pública*. [S.l.: s.n.], 2010. Citado na página 17.

MEYER, D. et al. *Package 'e1071'. Misc Functions of the Department of Statistics, Probability, Theory Group*. 2015. Citado 3 vezes nas páginas 115, 117 e 118.

MEYER, D.; WIEN, F. T. Support vector machines. *R News*, v. 1, n. 3, p. 23–26, 2001. Citado 2 vezes nas páginas 117 e 118.

MOLINARI, A. H.; TACLA, C. A. *Titulação Automática de Acórdãos Baseado em Ontologia Jurisprudencial*. 2010. Citado 2 vezes nas páginas 49 e 52.

MOORE, D. F. *Applied survival analysis using R*. [S.l.]: Springer, 2016. Citado na página 99.

MOTA, R. G. S. *Métodos de tratamento adequados de conflitos no poder judiciário*. 51 p. Dissertação (Monografia. (Especialização em Direito Processual Civil e Gestão de Processos)) — Escola Superior da Magistratura do Estado do Ceará, Fortaleza, 2014. Citado na página 38.

NETO, N. W. *Gestão de gabinetes de magistrados nas Câmaras Cíveis do Tribunal de Justiça do Rio Grande do Sul*. 209 p. Dissertação (Dissertação de Mestrado. (Mestrado Profissional em Poder Judiciário)) — Fundação Getúlio Vargas, Rio de Janeiro, 2010. Citado na página 20.

NOGUEIRA, E. G. *Sistema de gestão de unidade judicial*. 108 p. Dissertação (Dissertação de Mestrado. (Mestrado Profissional em Poder Judiciário)) — Fundação Getúlio Vargas, Rio de Janeiro, 2010. Citado na página 20.

NOGUEIRA, J. M.; PACHECO, R. S. *A gestão do poder judiciário nos estudos de administração pública*. 2009. 21 p. Citado na página 28.

NORTH, M. *Data mining for the masses*. CreateSpace Independent Publishing Platform: [s.n.], 2012. 264 p. Citado 7 vezes nas páginas 67, 93, 99, 103, 109, 114 e 119.

OLIVIERI, R. d. C. *Autos eletrônicos na Justiça Federal da 2ª Região: a contribuição do processo eletrônico na redução do tempo de tramitação dos processos*. 91 p. Dissertação (Dissertação de Mestrado. (Mestrado Profissional em Poder Judiciário)) — Fundação Getúlio Vargas, Rio de Janeiro, 2010. Citado na página 20.

PALVARINI, B. Guia referencial de mensuração do desempenho na administração pública. 2010. Citado na página 16.

PAVANELLI, A. M. *Utilização de redes neurais artificiais para a previsão do tempo de duração de audiencias trabalhistas*. 141 p. Dissertação (Dissertação de Mestrado. (Mestrado em Métodos Numéricos em Engenharia)) — Universidade Federal do Paraná, Curitiba, 2008. Citado 6 vezes nas páginas 23, 29, 49, 53, 54 e 57.

PAVANELLI, A. M. et al. Técnicas de reconhecimento de padrões aplicadas à justiça do trabalho. *Pesquisa Operacional para o Desenvolvimento*, Rio de Janeiro, v. 3, n. 2, p. 90–106, 2011. Citado 9 vezes nas páginas 22, 28, 29, 49, 51, 52, 53, 54 e 58.

PAVANELLI, G. *Análise do tempo de duração de processos trabalhistas utilizando redes neurais artificiais como apoio a tomada de decisões*. 132 p. Dissertação (Dissertação de Mestrado. (Mestrado em Métodos Numéricos em Engenharia)) — Universidade Federal do Paraná, Curitiba, 2007. Citado 7 vezes nas páginas 22, 23, 29, 49, 53, 54 e 56.

PAVANELLI, G.; PAVANELLI, A. M.; COSTA, D. M. B. Redes neurais artificiais, regressão linear múltipla e árvores de decisão aplicadas à previsão junto à justiça trabalhista. *XLV Sbpo-Simpósio Brasileiro de Pesquisa Operacional. Anais... Natal (RN): XLV SBPO*, p. 2322–2332, 2013. Citado 8 vezes nas páginas 22, 29, 49, 52, 53, 54, 59 e 102.

PEREIRA, M. O mau funcionamento do poder judiciário como empecilho ao desenvolvimento econômico brasileiro. *Constituição, economia e desenvolvimento: revista da Academia Brasileira de Direito Constitucional. Curitiba*, Curitiba, n. 2, p. 52–85, 2010. Citado na página 27.

PINHEIRO, V. d. A. *Poder judiciário: crise e reforma*. 81 p. Dissertação (Monografia. (Especialização em Direito Processual Civil e Gestão de Processos)) — Escola Superior da Magistratura do Estado do Ceará, Fortaleza, 2008. Citado na página 44.

PROVOST, F.; FAWCETT, T. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. Cambridge: O'Reilly Media, Inc., 2013. 409 p. Citado 8 vezes nas páginas 54, 64, 66, 99, 103, 110, 119 e 121.

PUCRS. *Demandas judiciais e morosidade da justiça civil*. Porto Alegre, 2011. Citado 2 vezes nas páginas 37 e 38.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>. Citado na página 112.

RAJARAMAN, A. et al. *Mining of massive datasets*. Cambridge: Cambridge University Press, 2012. 513 p. Citado 7 vezes nas páginas 64, 65, 66, 70, 73, 93 e 94.

RONG, X. Deepnet: deep learning toolkit in r. 2014. Citado 2 vezes nas páginas 115 e 117.

ROSA, F. I. et al. *Avaliação do tempo de tramitação dos processos em juizados especiais cíveis nas comarcas de Santa Catarina*. Dissertação (Dissertação de Mestrado. (Métodos e Gestão em Avaliação - PPGMGA)) — Unversidade Federal de Santa Catarina, Florianópolis, 2017. Citado 2 vezes nas páginas 22 e 60.

RUSCHEL, A. J. *Modelo de conhecimento para apoio ao juiz na fase processual trabalhista*. 206 p. Dissertação (Tese. (Doutorado em Engenharia e Gestão do Conhecimento)) — Universidade Federal de Santa Catarina, Florianópolis, 2012. Citado 4 vezes nas páginas 33, 34, 49 e 53.

SADEK, M. T. Judiciario: mudancas e reformas. *Estudos Avançados*, Scielo, v. 18, p. 79 – 101, 08 2004. ISSN 0103-4014. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142004000200005&nrm=iso>. Citado 4 vezes nas páginas 14, 36, 37 e 38.

SAUNDERS, M.; LEWIS, P.; THORNHILL, A. *Research methods for business students*. Fifth. Harlow: Prentice Hall, 2009. 649 p. Citado 3 vezes nas páginas 97, 98 e 99.

SCHEFFER, T. Finding association rules that trade support optimally against confidence. In: SPRINGER. *European conference on principles of data mining and knowledge discovery*. [S.l.], 2001. p. 424–435. Citado na página 94.

SCHNEIDER, L. F. *A aplicação do processo de descoberta de conhecimento em dados do poder judiciário do Estado do Rio Grande do Sul*. 103 p. Dissertação (Dissertação. (Mestrado em Informática)) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003. Citado 7 vezes nas páginas 22, 23, 29, 49, 53, 54 e 56.

SCHÖLKOPF, B. et al. Shrinking the tube: a new support vector regression algorithm. In: *Advances in neural information processing systems*. [S.l.: s.n.], 1999. p. 330–336. Citado 3 vezes nas páginas 76, 77 e 78.

SERBENA, C. A. Interfaces atuais entre a e-justiça e a q-justiça no brasil. *Revista de Sociologia e Política*, Universidade Federal do Paraná, v. 21, n. 45, p. 47, 2013. Citado 2 vezes nas páginas 39 e 44.

SHAMIR, N.; SHAMIR, J. The role of prosecutor's incentives in creating congestion in criminal courts. *Review of Law & Economics*, De Gruyter, v. 8, n. 3, p. 579–618, 2012. Citado 2 vezes nas páginas 19 e 22.

SILVA, E. L. da; MENEZES, E. M. *Metodologia da pesquisa e elaboração de dissertação*. fourth. Florianópolis: Universidade Federal de Santa Catarina, 2005. 139 p. Citado na página 97.

SILVA, J. M. C. d. *O processo eletrônico frente aos princípios da celeridade processual e do princípio doacesso à justiça*. 123 p. Dissertação (Dissertação de Mestrado. (Mestrado em Direito) — Universidade Católica de Pernambuco, Recife, 2015. Citado 3 vezes nas páginas 36, 37 e 38.

SILVA, W. F.; CUNHA, J.; TALON, A. F. Mineração de dados: análise de duração de processos jurídicos do estado de São Paulo. *Revista Eletrônica e-Fatec*, v. 3, n. 1, 2013. Citado 7 vezes nas páginas 22, 23, 29, 49, 52, 54 e 58.

SILWATTANANUSARN, T.; TUAMSUK, K. *Data mining and its applications for knowledge management: a literature review from 2007 to 2012*. 2012. 13–24 p. Citado 2 vezes nas páginas 49 e 64.

SPURR, S. J. The duration of personal injury litigation. In: *Research in Law and Economics*. [S.l.]: Emerald Group Publishing Limited, 2000. p. 223–246. Citado na página 21.

TRF4. *Competência e Organização do TRF da 4ª Região*. Porto Alegre: [s.n.], 2018. Disponível em <http://www2.trf4.jus.br/trf4/controlador.php?acao=pagina_visualizarid_pagina=1>. Acesso em 21 mai. 2018. Citado 2 vezes nas páginas 35 e 36.

TRIVIÑOS, A. N. S. *Introdução à pesquisa em ciências sociais: a pesquisa qualitativa em educação. São Paulo: Atlas, 1987*. São Paulo: Atlas, 1987. 175 p. Citado na página 98.

TUBINO, D. F. *Sistemas de produção: a produtividade no chão de fábrica*. Porto Alegre: Bookman, 2004. 134 p. Citado 2 vezes nas páginas 47 e 48.

WALSH, W. A. et al. Length of time to resolve criminal charges of child sexual abuse: A three-county case study. *Behavioral sciences & the law*, Wiley Online Library, v. 33, n. 4, p. 528–545, 2015. Citado na página 19.

YIN, R. K. *Estudo de Caso: Planejamento e Métodos*. Second. Porto Alegre: Bookman editora, 2004. 205 p. Citado na página 98.

ZAKI, M. J.; JR, W. M. *Data mining and analysis: fundamental concepts and algorithms*. First. New York: Cambridge University Press, 2006. 604 p. Citado 8 vezes nas páginas 65, 70, 72, 73, 93, 94, 96 e 119.

ZHANG, H. et al. Class association rule mining with multiple imbalanced attributes. In: SPRINGER. *Australasian Joint Conference on Artificial Intelligence*. [S.l.], 2007. p. 827–831. Citado na página 94.

ZHOU, J. D. Determinants of delay in litigation: Evidence and theory. In: BEPRESS. *American Law & Economics Association Annual Meetings*. [S.l.], 2008. p. 21. Citado 2 vezes nas páginas 21 e 23.

ZUMEL, N.; MOUNT, J.; PORZAK, J. *Practical data science with R*. New York: Manning, 2014. 389 p. Citado 9 vezes nas páginas 62, 63, 71, 76, 93, 94, 96, 119 e 120.

Apêndices

APÊNDICE A – FORMULÁRIO PARA ENTREVISTAS COM ESPECIALISTAS

Influência das variáveis no tempo de tramitação processual de 2º grau

1. Email address *

As perguntas seguintes têm o objetivo de identificar a influência de variáveis no tempo de tramitação processual de 2º Grau. Assinale o grau de importância que, na sua opinião, cada variável constante no Business Intelligence têm sobre o tempo de tramitação processual.

2. Classe

	1	2	3	4	5	6	7	
Pouco Importante	◯	◯	◯	◯	◯	◯	◯	Muito Importante

3. Assunto *

	1	2	3	4	5	6	7	
Pouco Importante	◯	◯	◯	◯	◯	◯	◯	Muito Importante

4. Justiça de origem (Federal, Estadual, etc.)

	1	2	3	4	5	6	7	
Pouco Importante	◯	◯	◯	◯	◯	◯	◯	Muito Importante

5. Unidade da Federação de Origem

	1	2	3	4	5	6	7	
Pouco Importante	◯	◯	◯	◯	◯	◯	◯	Muito Importante

6. Meio processual

	1	2	3	4	5	6	7	
Pouco Importante	◯	◯	◯	◯	◯	◯	◯	Muito Importante

	1	2	3	4	5	6	7	
Pouco Importante	◯	◯	◯	◯	◯	◯	◯	Muito Importante

7. Gabinete (de distribuição)

8. Entidade *

	1	2	3	4	5	6	7	
Pouco Importante	◯	◯	◯	◯	◯	◯	◯	Muito Importante

9. Na sua opinião, de forma geral, qual o nível do assunto, de acordo com a TUA - Tabela Única de Assuntos da Justiça Federal, melhor descreve as ações:

	1	2	3	4	5	6	7	
Pouco Importante	◯	◯	◯	◯	◯	◯	◯	Muito Importante

( ) Nível 1. Por Exemplo: Direito Civil
( ) Nível 2. Por Exemplo: Coisas
( ) Nível 3. Por Exemplo: Propriedade
( ) Nível 4. Por Exemplo: Aquisição
( ) Nível 5. Por exemplo: Usucapião da Lei nº 6969/1981

10. O espaço a seguir está destinado a comentarios sobre a influência da CLASSE no tempo de tramitação do processo.

11. O espaço a seguir está destinado a comentarios sobre a influência da ASSUNTO no tempo de tramitação do processo.

12. O espaço a seguir está destinado a comentários sobre a influência da ORIGEM DA JUSTIÇA no tempo de tramitação do processo.

13. O espaço a seguir está destinado a comentários sobre a influência da UNIDADE DA FEDERAÇÃO DE ORIGEM no tempo de tramitação do processo.

14. O espaço a seguir está destinado a comentários sobre a influência do MEIO (FÍSICO/ELETRÔNICO) no tempo de tramitação do processo.

15. O espaço a seguir está destinado a comentários sobre a influência do GABINETE (de distribuição) no tempo de tramitação do processo.

16. O espaço a seguir está destinado a comentários sobre a influência da ENTIDADE no tempo de tramitação do processo.

17. O espaço a seguir está destinado a comentários sobre o que influencia o tempo de tramitação processual.

APÊNDICE B – LEGENDA DAS VARIÁVEIS

1. Justiça de Origem do processos:
   JE - Justiça Estadual
   JF - Justiça Federal
   TRF - Tribunal Regional Federal
   Od – Outras origens

2. Competência do processo:
   Ad – Administrativo
   Pr – Previdenciário
   Tr – Tributário Demais - Demais competências

3. Classe processual:
   C10828 - Apelação Cível
   C10992 - Apelação / Reexame Necessário
   C10822 - Agravo de Instrumento
   C10855 - Reexame Necessário Cível
   C11011 - Ação Rescisória - Seção
   C10943 - Ação Rescisória
   C10977 - Embargos Infringentes
   C11009 - Mandado de Segurança - Turma
   C40012 - Pedido de Efeito Suspensivo á Apelação – Turma
   CDemais – Demais classes

4. Assunto:
   040119 - Aposentadoria por Tempo de Contribuição (Art. 55/6)
   N0312 - Dívida Ativa
   N040310 - Renúncia ao benefício
   N040105 - Auxílio-Doença Previdenciário
   N040101 - Aposentadoria por Invalidez (Art. 42/7)
   N040104 - Aposentadoria Especial (Art. 57/8)
   N04010202 - Aposentadoria por Idade - Rural (art. 48/51)
   N040118 - Aposentadoria por Tempo de Serviço (Art. 52/4)
   N040102 - Aposentadoria por Idade (Art. 48/51)
   N04020113 - IRSM de Fevereiro de 1994(39,67%)

N06040101 - Expurgos Inflacionários / Planos Econômicos
N01040411 - Fornecimento de Medicamentos
N040103 - Aposentadoria por Tempo de Serviço (Art. 52/6) e/ou Tempo de Contribuição
N040203 - Reajustes e Revisões Específicos
N01031001 - Multas e demais Sanções
N010808 - Seguro-desemprego
N011501 - Multas e demais Sanções
N04020116 - Alteração do coeficiente de cálculo do benefício
N040501 - Averbação/Cômputo/Conversão de tempo de serviço especial
N040201 - RMI - Renda Mensal Inicial
N040107 - Salário-Maternidade (Art. 71/73)
N0219033205 - Seguro
N0402 - RMI - Renda Mensal Inicial, Reajustes e Revisões Específicas
N040113 - Benefício Assistencial (Art. 203,V CF/88)
N04020108 - Limitação do salário-de-benefício e da renda mensal inicial
N0401 - Benefícios em Espécie
N040404 - Concessão
N04011301 - Pessoas com deficiência
N02190405 - Cédula de crédito rural
N0115 - Dívida Ativa não-tributária
N040502 - Averbação/Cômputo de tempo de serviço de segurado especial (regime de economia familiar)
N030201 - IRPF/Imposto de Renda de Pessoa Física
N03040202 - Cofins
N040111 - Auxílio-Acidente (Art. 86)
N02190338 - Contratos Bancários)
NDemais – Demais assuntos

5. Meio processual:
   FI - Físico
   Ele - Eletrônico

6. Unidade da federação:
   PR - Paraná RS - Rio Grande do Sul
   SC - Santa Catarina

7. Entidade:
   AGU-A: Advocacia Geral da União (Autora)

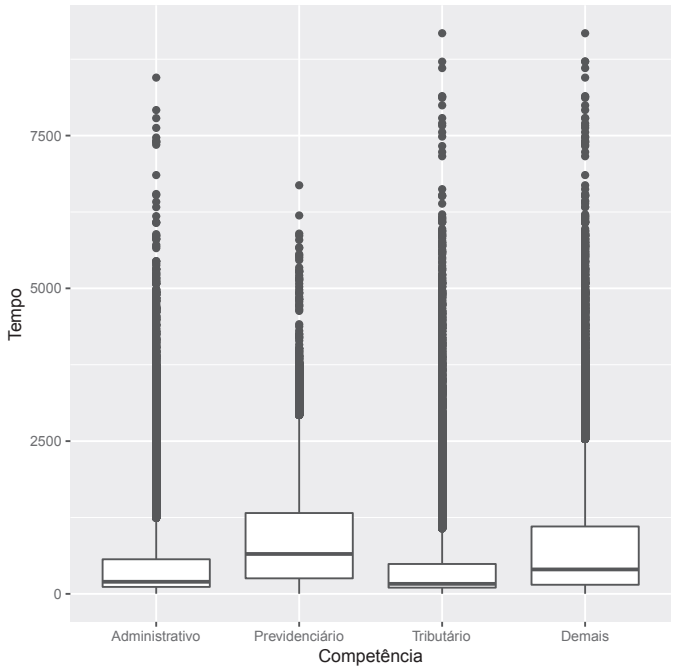
AGU-R: Advocacia Geral da União (Ré)
ANTT-A: Agência Nacional de Transportes Terrestres (Autora)
ANTT-A: Agência Nacional de Transportes Terrestres (Ré)
Banco-A: Bancos (Autores)
Banco-A: Bancos (Réus)
CEF-A: Caixa Econômica Federal (Autora)
CEF-R: Caixa Econômica Federal (Ré)
Conselho-A: Conselhos Regionais e Federais (Autores)
Conselho-R: Conselhos Regionais e Federais (Réus)
DNIT-A: Departamento Nacional de Infraestrutura de Transportes(Autor)
DNIT-R: Departamento Nacional de Infraestrutura de Transportes (Réu)
ELETROBRAS-A: Centrais Elétricas Brasileiras (Autora)
ELETROBRAS-Ré: Centrais Elétricas Brasileiras (Ré)
EMGEA-A: Empresa Gestora de Ativos (Autora)
EMGEA-R: Empresa Gestora de Ativos (Ré)
Estado-A: Estados (PR, RS, SC) (Autores)
Estado-R: Estados (PR, RS, SC) (Réus)
Fazenda Nacional-A: União – Fazenda Nacional (Autora)
Fazenda Nacional-R: União – Fazenda Nacional (Ré)
FNDE-A: Fundo Nacional de Desenvolvimento da Educação (Autor)
FNDE-A: Fundo Nacional de Desenvolvimento da Educação (Réu)
IBAMA-A: Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (Autor)
IBAMA-R: Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (Réu)
IF-A: Institutos Federais (Autores)
IF-R: Institutos Federais (Réus)
INCRA-A: Instituto Nacional de Colonização e Reforma Agrária (Autor)
INCRA-R: Instituto Nacional de Colonização e Reforma Agrária (Réu)
INMETRO-A: Instituto Nacional de Metrologia, Qualidade e Tecnologia (Autor)
INMETRO-R: Instituto Nacional de Metrologia, Qualidade e Tecnologia (Réu)
INSS-A: Instituto Nacional de Seguridade Social (Autor)
INSS-R: Instituto Nacional de Seguridade Social (Réu)
MPF-A: Ministério Público Federal (Autor)
MPF-R: Ministério Público Federal (Réu)
MUNICIPIO-A: Municípios (Porto Alegre, Alvorada, etc.) (Autores)
MUNICIPIO-R: Municípios (Porto Alegre, Alvorada, etc.) (Réus)
União Federal-A: União Federal (Autora)
União Federal-R: União Federal (Ré)

Universidade-A: Universidades Federais (Autoras)
Universidade-R: Universidades Federais (Rés)

APÊNDICE C – ANÁLISE DESCRITIVA DO TEMPO DE ATRAVESSAMENTO POR VARIÁVEL CATEGÓRICA

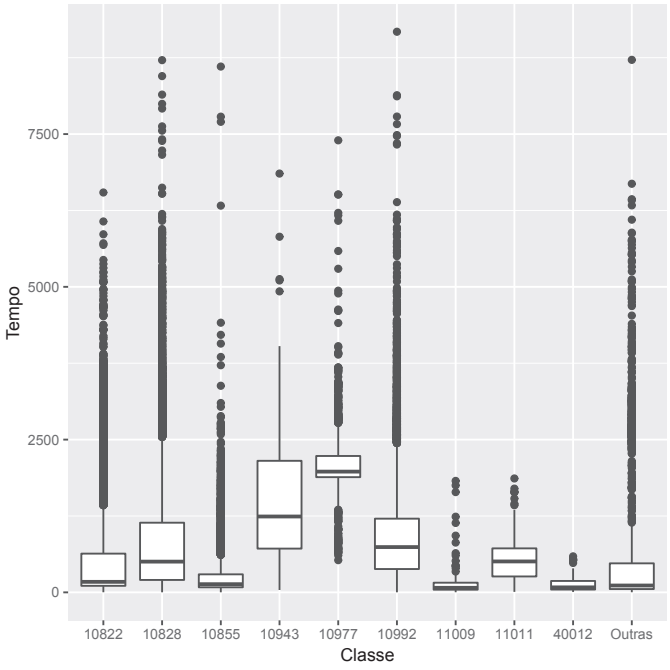
A seguir a comparação das categorias das variáveis quanto ao tempo de atravessamento mediante boxplots.

Figura 39 – Boxplot - competência



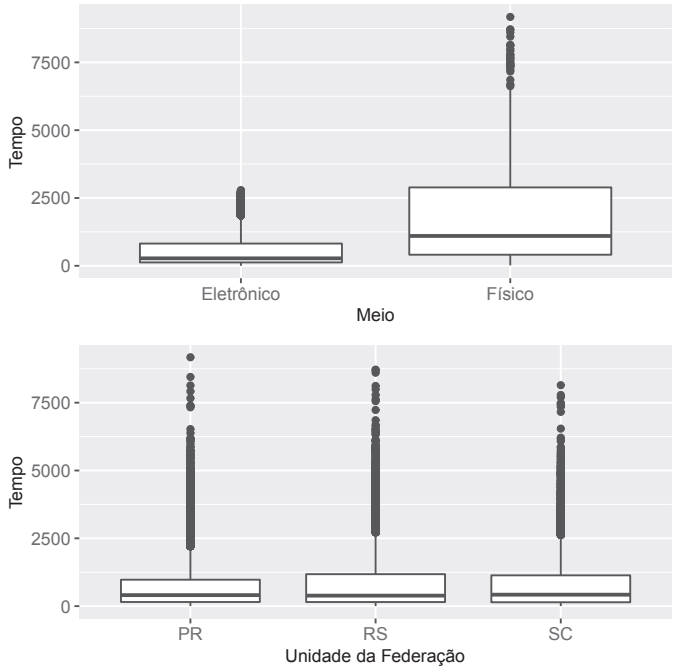
Fonte: Dados da pesquisa

Figura 40 – Boxplot - Classe



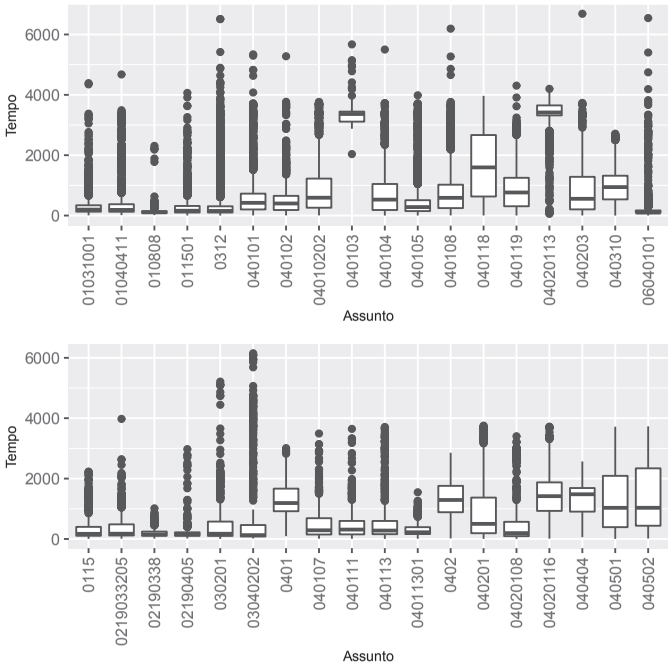
Fonte: Dados da pesquisa

Figura 41 – Boxplot - Unidade da federação e meio processual



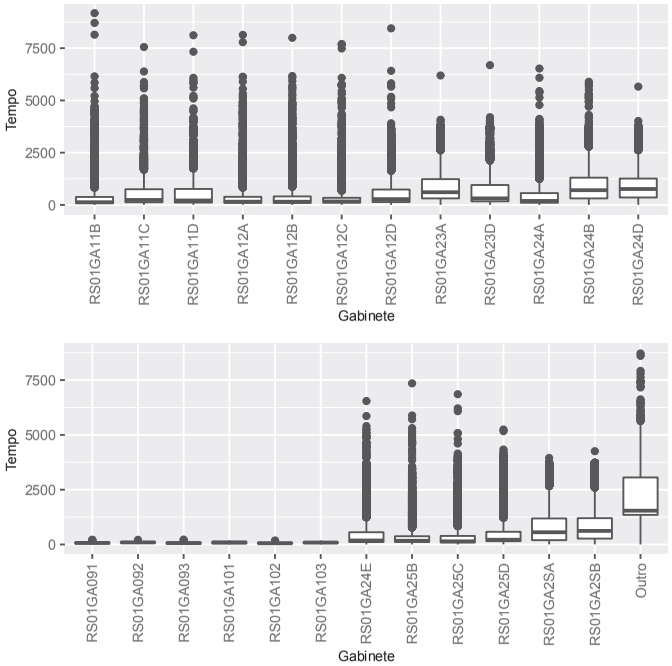
Fonte: Dados da pesquisa

Figura 42 – Boxplot - Assunto



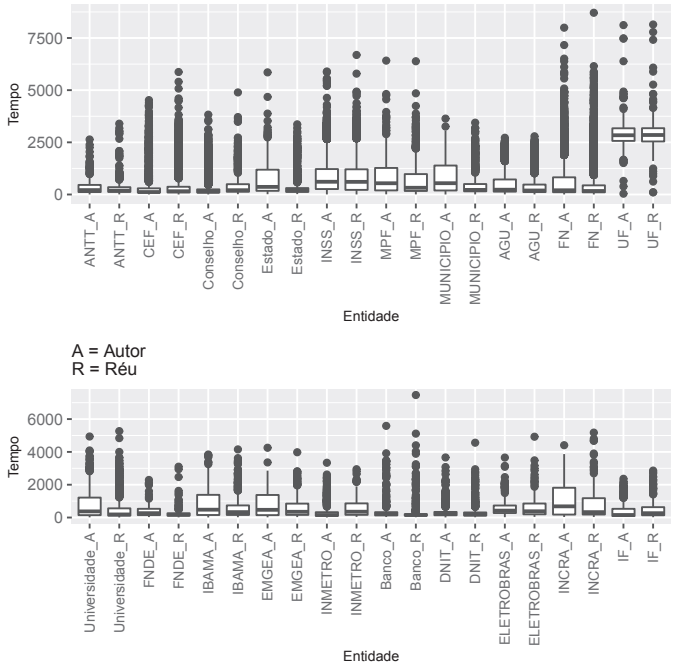
Fonte: Dados da pesquisa

Figura 43 – Boxplot - Gabinete



Fonte: Dados da pesquisa

Figura 44 – Boxplot - Entidade



Fonte: Dados da pesquisa

Tabela 22 – Medidas descritivas das principais covariáveis

Covariável	%	Mediano	Médio	Desvio Padrão
Ad	26,4	198	541,9	781,060
Pr	57,4	653	969,3	941,630
Tr	15,8	164	520,2	844,770
Demais	0,4	2634	2182,4	1530,280
RS	47,2	387	831,7	972,490
PR	28,1	406	724,1	827,480
SC	24,3	422	797,3	916,350
FI	23,7	1101	1587,9	1304,410
Ele	76,3	275	543,5	568,540
JE	26,7	407	628,3	649,960
JF	72,6	388	842	985,340
TRF	0,6	1792	1811,8	1441,120
Od	0,0	1092	1264,5	957,020
C10828	43,4	503	774,4	766,490
C10822	30,5	172	793,6	1193,180
C10992	19,1	741	885,4	691,000
C10855	4,1	134	279,1	430,660
C10977	0,8	1976	2136,3	649,420

Continua...

Tabela 22 – Medidas descritivas das principais covariáveis

Covariável	%	Mediano	Médio	Desvio padrão
C11011	0,2	506	551,5	378,860
C11009	0,1	74	155,9	264,920
C40012	0,1	82	132,1	123,930
C10943	0,1	1242	1531	1185,000
CDemais	1,5	112	755	1277,970
CEF (Autor)	2,4	118	393,5	676,520
CEF (Réur)	4,4	156	353,9	527,020
FNDE (Réur)	0,1	170	275,8	404,730
INCRA (Autor)	0,1	682	1063,5	1019,370
INCRA (Réur)	0,1	330	918,6	1160,300
IBAMA (Autor)	0,4	478	895,5	923,390
IBAMA (Réur)	0,4	330	623,6	726,040
UF (Autor)	0,2	2839	2905,8	845,060
UF (Réur)	0,2	2851	2891,5	904,190
INMETRO (Autor)	0,5	155	339,3	471,620
INMETRO (Réur)	0,1	355	612,1	612,500
IF (Autor)	0,2	155	419,4	512,960
IF (Réur)	0,2	264	519,5	573,220
Universidade (Autor)	0,7	375	756,5	826,290
Universidade (Réur)	0,8	188	547,2	764,650
Demais (Autor)	1,3	276	686,2	787,290
Demais (Réur)	1,6	230	549,9	696,720
N040119	8,8	764	839,5	615,550
N0312	8,6	153	364,8	566,260
N040310	7,6	943	962,8	546,060
N040105	6,4	281	425,3	493,700
N040101	4,2	422	581,9	607,730
N040104	4,2	528	719,5	667,160
N04010202	4,1	591	903,2	883,440
N040108	2,3	587	833,4	835,100
N040118	2,1	1596	1627,4	1035,350
N040102	1,8	400	508,7	498,440
N04020113	1,6	3416	3367,2	469,770
Conselho (Autor)	4,0	133	302,9	437,99
Conselho (Réur)	0,6	206	452,4	584,81
Estado (Autor)	0,5	365	814,8	867,59

Continua...

Tabela 22 – Medidas descritivas das principais covariáveis

Covariável	%	Mediano	Médio	Desvio padrão
Estado (Réur)	0,6	187	392,5	560,57
INSS (Autor)	29,1	610	883,0	814,06
INSS (Réur)	26,2	606	799,3	697,22
MPF (Autor)	0,3	539	944,1	984,57
MPF (Réur)	0,5	322	734,2	852,95
MUNICIPIO (Autor)	0,4	542	836,3	805,70
MUNICIPIO (Réur)	1,1	220	423,3	502,43
AGU (Autor)	3,1	230	549,2	633,16
AGU (Réur)	4,0	192	423,6	525,41
ANTT (Autor)	0,2	211	376,0	421,86
ANTT (Réur)	0,3	183	365,0	516,49
Banco (Autor)	0,3	210	420,4	708,82
Banco (Réur)	1,0	154	235,1	431,30
DNIT (Autor)	0,5	215	338,3	400,30
DNIT (Réur)	0,4	186	354,6	512,59
ELETROBRAS (Autor)	0,3	420	591,6	543,17
ELETROBRAS (Réur)	0,1	381	681,1	778,76
EMGEA (Autor)	0,1	470	845,3	861,88
EMGEA (Réur)	0,1	351	682,9	765,23
FN (Autor)	5,1	202	632,4	840,60
FN (Réur)	7,0	168	466,1	719,71
FNDE (Autor)	0,1	266	435,2	461,74
N06040101	1,6	98	218,6	466,66
N01040411	1,3	187	487,9	714,61
N040103	1,3	3365	3341,8	253,76
N040203	1,0	555	826,5	760,91
N01031001	0,9	198	349,4	464,48
N010808	0,9	109	131,5	151,94
N011501	0,9	160	340,6	512,50
N04020116	0,8	1418	1478,4	820,27
N040501	0,8	1032	1316,3	1043,25
N040201	0,7	500	956,4	1028,45
N040107	0,7	287	499,8	512,45
N0219033205	0,7	174	360,0	397,88
N0402	0,7	1292	1310,2	589,86
N040113	0,7	279	562,4	743,06

Continua...

Tabela 22 – Medidas descritivas das principais covariáveis

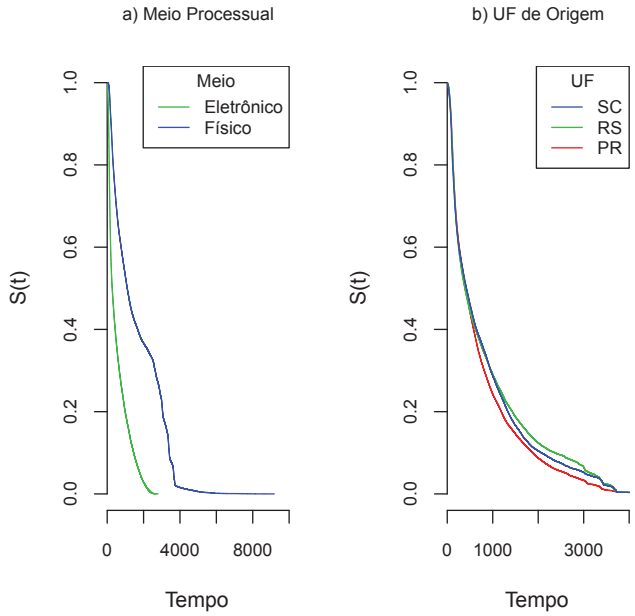
Covariável	%	Mediano	Médio	Desvio padrão
N04020108	0,6	194	525,9	755,83
N0401	0,6	1190	1284,1	506,18
N040404	0,5	1481	1316,9	491,49
N04011301	0,5	222	312,9	238,46
N02190405	0,5	146	224,6	334,96
N0115	0,5	170	345,4	426,72
N040502	0,5	1032	1379,9	1081,81
N030201	0,5	173	529,0	823,84
N03040202	0,5	127	866,8	1488,21
N040111	0,5	311	488,2	537,97
N02190338	0,4	150	197,2	146,86
NDemais	29,3	265	723,7	947,58

Fonte: Dados da pesquisa

APÊNDICE D – FUNÇÕES DE SOBREVIVÊNCIA POR VARIÁVEL CATEGÓRICA

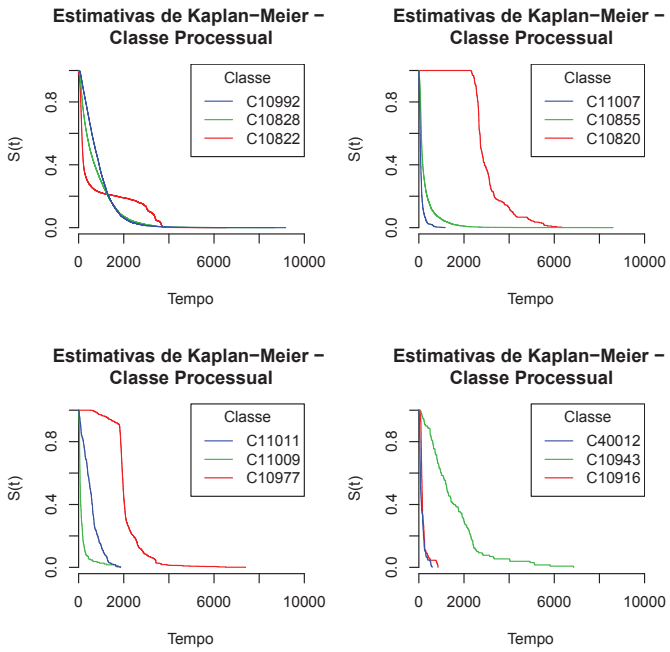
As figuras de 45 a 52 mostram as estimativas de soobrevivência da função de Kaplan Meyer de acordo com as covariáveis.

Figura 45 – Estimativas de Kaplan Meyer para meio processual e origem



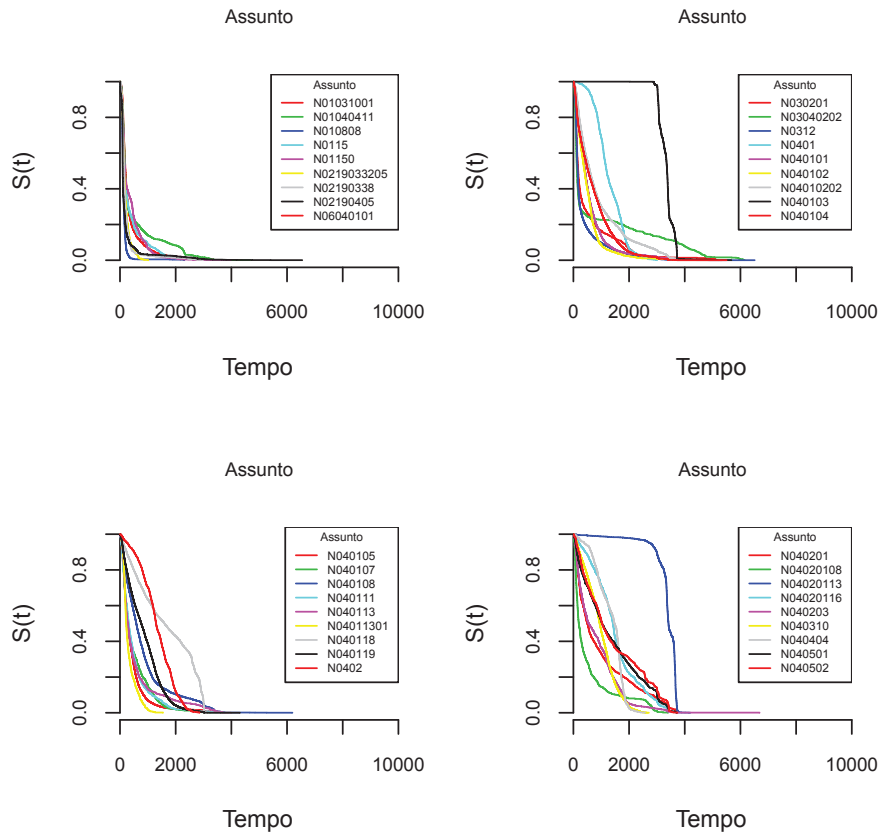
Fonte: Dados da pesquisa

Figura 46 – Estimativas de Kaplan Meyer para classe



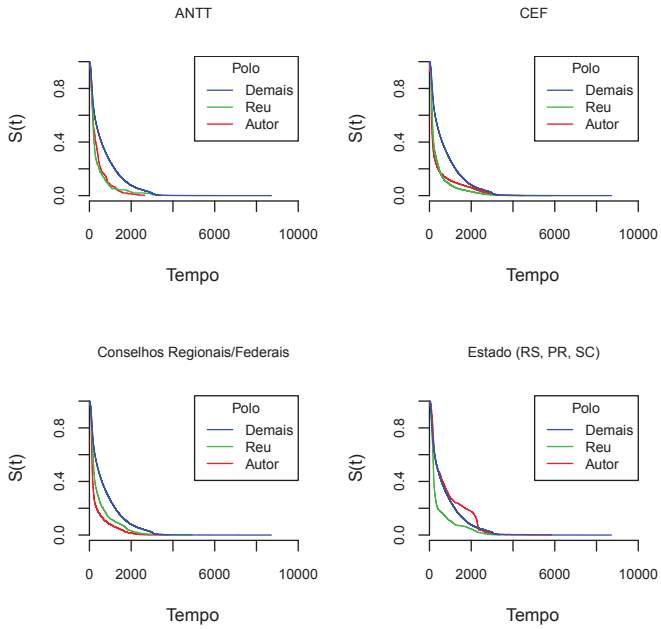
Fonte: Dados da pesquisa

Figura 47 – Estimativas de Kaplan Meyer para assunto



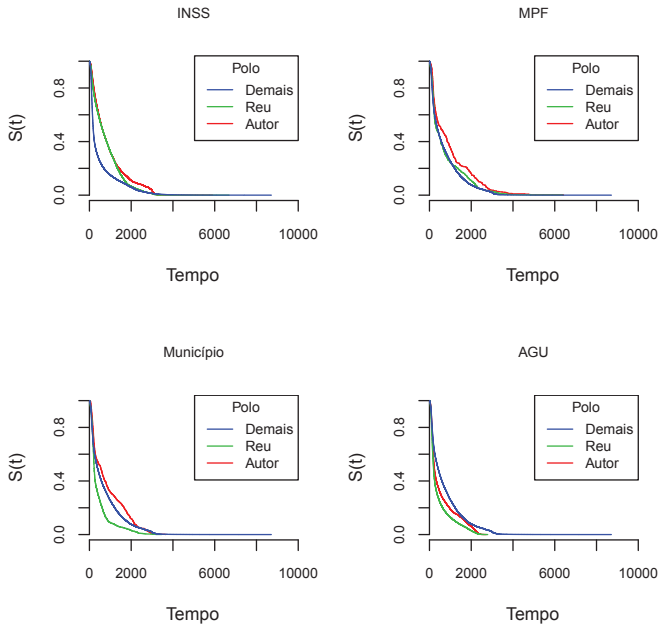
Fonte: Dados da pesquisa

Figura 48 – Estimativas de Kaplan Meyer para entidade



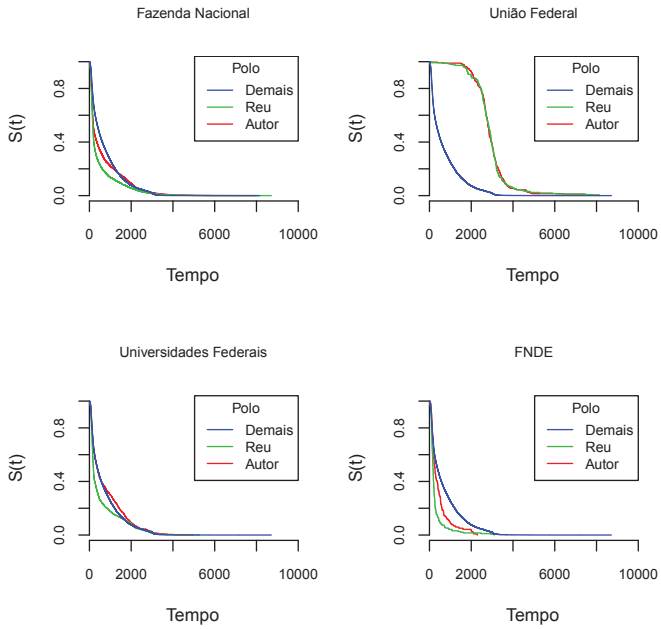
Fonte: Dados da pesquisa

Figura 49 – Estimativas de Kaplan Meyer para entidade - continuação



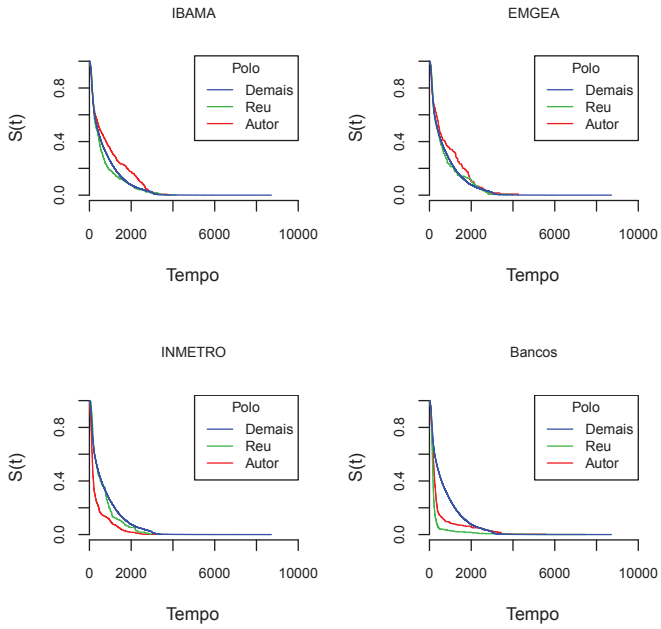
Fonte: Dados da pesquisa

Figura 50 – Estimativas de Kaplan Meyer para entidade - continuação



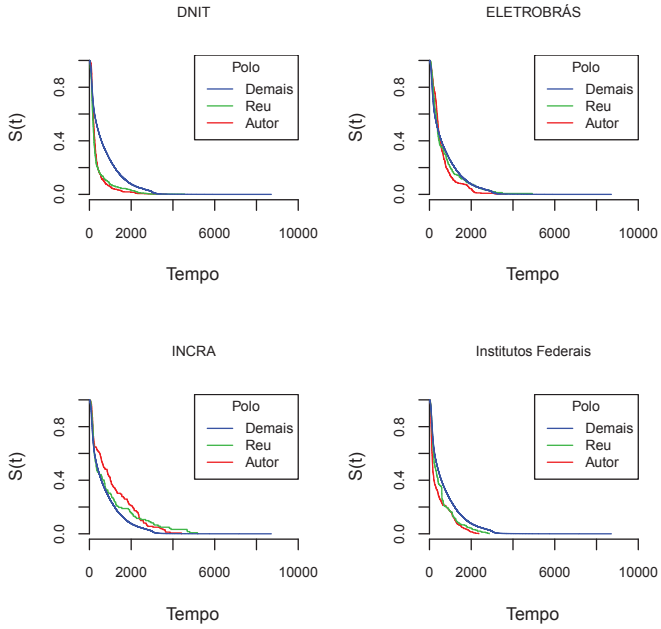
Fonte: Dados da pesquisa

Figura 51 – Estimativas de Kaplan Meyer para entidade - continuação



Fonte: Dados da pesquisa

Figura 52 – Estimativas de Kaplan Meyer para entidade - continuação



Fonte: Dados da pesquisa

APÊNDICE E – ANÁLISE DE COMPONENTES PRINCIPAIS

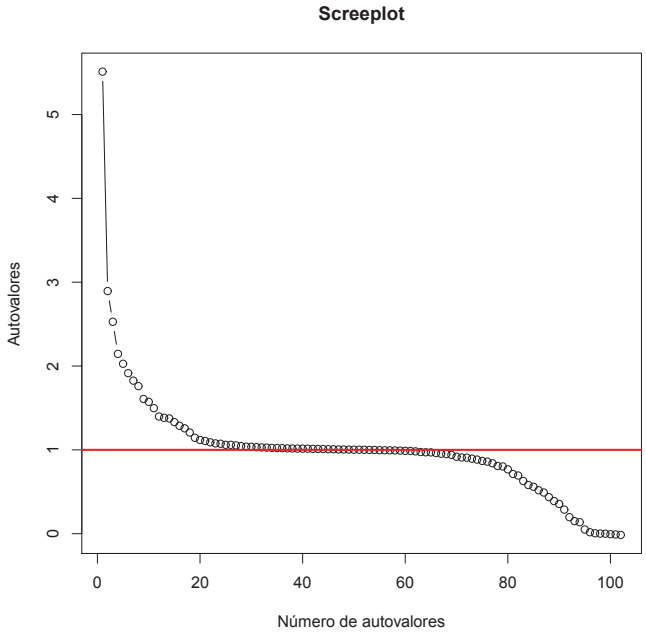
A Análise de componentes principais (ACP) é uma técnica de análise multivariada a qual transforma um conjunto de k variáveis correlacionadas em um conjunto de k componentes não correlacionados, resultado de combinação linear das variáveis originais. Os primeiros M componentes concentram a maior parte da variância dos dados.

Com o objetivo de auxiliar na determinação do número de componentes principais que poderiam ser usados, caso a matriz de dados original fosse substituída por um conjunto menor, foi construído um *screeplot*. *Screeplot* é um gráfico que mostra os autovalores de uma matriz. Em ACP, o objetivo é obter o número de componentes seguindo alguns critérios:

- de acordo com algum percentual da variância;
- número de autovalores maiores que a unidade;
- de acordo com a declividade do gráfico *screeplot*;

O gráfico da figura 53 mostra o *screeplot* para os dados da análise. Assim, em vez das covariáveis originais, seriam usados um número menor de componentes. Segundo o gráfico, e de acordo com o critério do percentual de variância mínima, 65 componentes explicam 80% da variância original. Conforme o critério de autovalores maiores que 1, seriam 52 componentes explicando 67% da variância.

Figura 53 – Análise de Componentes Principais



Fonte: Dados da pesquisa