Diagramas de influência: uma aplicação em Jurimetria

Julio Adolfo Zucon Trecenti

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestrado em Estatística

Programa: Estatística

Orientador: Prof. Dr. Carlos Alberto de Bragança Pereira

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CNPq.

São Paulo, novembro de 2015

Diagramas de influência:
uma aplicação em Jurimetria

Esta é a versão original da dissertação elaborada pelo candidato (Julio Adolfo Zucon Trecenti), tal como submetida à Comissão Julgadora.

Resumo

TRECENTI, J. A. Z. **Diagramas de influência: uma aplicação em Jurimetria**. 2015. 120 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

Esse trabalho contribui para o desenvolvimento da jurimetria, uma disciplina do direito que utiliza a estatística na elucidação de fenômenos jurídicos. Motivados pela flexibilidade e facilidade de interpretação, desenvolvemos um diagrama de influência aplicado a uma base de dados de processos cíveis. A base utilizada foi obtida diretamente da web, por meio de técnicas de raspagem de dados e análise de textos. Tanto a extração dos dados quanto a modelagem foram incorporados a pacotes do R, possibilitando novas pesquisas com diferentes diagramas e bases de dados. Nosso modelo final foi utilizado para predizer as decisões dos processos de acordo com diferentes níveis de informação. Os resultados foram equivalentes a modelos de florestas aleatórias. O modelo também foi utilizado para decidir se seria mais vantajoso entrar com um processo na Justiça Comum ou nos Juizados Especiais Cíveis, considerando diferentes cenários de conflitos com empresas.

Palavras-chave: jurimetria, diagramas de influência, redes Bayesianas, inferência Bayesiana, dados com omissão, raspagem de dados da web, mineração de texto, direito civil.

Abstract

TRECENTI, J. A. Z. **Influence diagrams: application in Jurimetrics**. 2015. 120 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

This work makes some contribution to jurimetrics, a field of study that uses statistics to elucidade legal phenomena. We develop an influence diagram that applies to civil justice litigation and that is flexible and easy to interpret. All our data were obtained on the web, through web scraping and text mining techniques. We also develop some R packages to load the data and model hybrid Bayesian networks with missing data. Finally, we combine the output of these packages with decision theory to decide under what circumstances it is better to litigate on small claims justice than on standard courts.

Keywords: jurimetrics, influence diagrams, Bayesian networks, Bayesian analysis, incomplete data, web scraping, text mining, civil rights.

Sumário

Capítulo 1

Introdução

Abstração e concretude. Determinismo e aleatório. Direito e estatística. Uma rima rara. Assim é a jurimetria.

A jurimetria, que por definição é a utilização da estatística no direito, é uma disciplina que procura investigar fenômenos jurídicos através da observação empírica. Por situar o objeto de estudo no tempo e no espaço, a jurimetria funciona como uma abordagem complementar para solucionar os diversos questionamentos presentes no direito.

A jurimetria ainda vive sua infância e não tem limites bem definidos. As análises, com poucas diretrizes, muitas vezes limitam-se a estudos descritivos, ou então importam métodos já consolidados de outras áreas do conhecimento, sem considerarem as suposições dos modelos.

Nesse trabalho, buscamos levar a pesquisa em jurimetria a um novo patamar. Discutimos alguns conceitos básicos e descrevemos uma modelagem baseada em *diagramas de influência*, que é adaptável a diversas situações e problemas.

Para ilustrar a teoria desenvolvida, utilizamos como base de dados as sentenças de 2014 de processos cíveis envolvendo empresas de grande porte no Tribunal de Justiça de São Paulo (TJSP). O tema foi escolhido pela elevada quantidade de litígios e relevância desses conflitos, tanto para as pessoas que se sentem lesadas, quanto para empresas que buscam estratégias para minimizar perdas.

No estudo, construímos um diagrama de influência capaz de predizer com acurácia razoável as decisões dos processos com base em informações parciais. Além disso, o modelo sugere se o autor do processo deveria iniciar o litígio na Justiça Comum ou nos Juizados Especiais Cíveis (JECs), considerando diferentes cenários.

No decorrer das atividades procuramos discutir a jurimetria de forma simples e pragmática. Na aplicação, buscamos tornar cada tarefa reprodutível, incluindo coleta dos dados, especificação, ajuste e validação de modelos.

Desenvolvemos a dissertação para que seja acessível a qualquer pesquisador interessado no tema, sem ser necessário conhecer técnicas estatísticas avançadas nem as regras do ordenamento jurídico. Pesquisadores que já trabalham com jurimetria poderão comparar as técnicas apresentadas com suas próprias experiências.

1.1 Motivação

A conveniência da utilização da jurimetria no estudo de fenômenos jurídicos pode ser explicada a partir de diversos pontos de vista. Em políticas públicas, por exemplo, a estatística auxilia na elaboração de estudos de impacto regulatório e sugestões de melhoria no sistema judiciário. Para administração dos tribunais, existem trabalhos como o Relatório Justiça em Números, do Conselho Nacional de Justiça[1], que visam descrever e mensurar o desempenho de tribunais com base em estatísticas de produtividade.

No mundo corporativo, a jurimetria é útil em diversas questões práticas. Muitas vezes, os problemas estão relacionados a gestão de grandes carteiras de processos, como a elaboração de estudos de risco e provisionamento, previsão das decisões dos juízes e desenvolvimento de estratégias advocatícias para processos de massa. A utilização da estatística algumas vezes é fundamental até mesmo para elaboração de provas em tipos de processos específicos, como perda de uma chance e lucros cessantes, por serem essencialmente problemas que envolvem incerteza.

Atualmente há amplo espaço para produção acadêmica em jurimetria. Nos Estados Unidos, por exemplo, um dos principais termos utilizados são os *Empirical Legal Studies* (ELS), e envolvendo pesquisas de diversas correntes do direito, economia, sociologia e psicologia, organizados no *Journal of Empirical Legal Studies* (JELS). Outra revista importante é o *Jurimetrics Journal*, que tem como prerrogativa a aplicação de ciência e tecnologia no direito. Podemos também encontrar textos relevantes sobre jurimetria nas revistas *Journal of Legal Studies*, *Journal of Law and Economics*, *Law, Probability and Risk* e *Journal of Law, Economics and Organization*.

No Brasil, temos poucos pesquisadores e institutos que se dedicam à jurimetria. O termo apareceu pela primeira vez em Ribeiro (1998), mas só é discutido com profundidade em Nunes (2012). O primeiro artigo sobre o tema escrito por estatísticos e profissionais do direito é de Zabala e Silveira (2014). A Rede de Estudos Empíricos no Direito (REED)[2] também publica anualmente uma revista de produção própria. O crescente interesse no tema e o surgimento de produção acadêmica sugerem boas oportunidades para pesquisadores interessados em jurimetria, tanto no Brasil quanto no exterior.

Em relação à modelagem estatística proposta, acreditamos que os diagramas de influência podem

[1] http://www.cnj.jus.br/programas-e-acoes/pj-justica-em-numeros.
[2] http://reedpesquisa.org/

estar sendo sub-utilizados pela comunidade científica. Em Pearl (2005), Judea Pearl discute os motivos pelos quais acredita que os diagramas de influência, da forma que foram definidos em Howard (1984), não foram prontamente implementados pelos pesquisadores da época, sendo gradativamente substituídos pela teoria de redes Bayesianas. Os modelos gráficos podem ser interpretados como formas de representar problemas complexos em sistemas que são, ao mesmo, tempo, i) razoáveis do ponto de vista teórico; ii) bem definidos, do ponto de vista estatístico; e iii) tratáveis do ponto de vista computacional. Quando um problema envolve decisões, a utilização de diagramas de influência ocorre naturalmente e permite a modelagem de decisões aparentemente complexas de forma simplificada.

Em relação à aplicação escolhida, a motivação para elaboração do estudo foi um problema ocorrido com este autor. Curiosamente, o meu nome foi inserido nos cadastros de inadimplentes pois eu supostamente não teria pago algumas mensalidades referentes à utilização de energia elétrica. Ocorre que nos meses referidos, eu já estava morando em outro lugar, e havia feito verbalmente o pedido de desligamento da luz no apartamento. Após utilizar o Serviço de Atendimento ao Cliente, chat on-line e deslocar-me até a empresa para registrar uma reclamação, não restaram alternativas a não ser entrar com uma ação judicial. Uma pergunta que me veio à mente foi: "devo entrar com ação em um Juizado Especial Cível ou na Justiça Comum?". Apresentaremos uma solução a essa questão com nossa aplicação.

1.2 Objetivos

Os objetivos que procuramos atingir nos próximos capítulos são:

1. Descrever e discutir os conceitos básicos de jurimetria e diagramas de influência.
2. A partir de uma base de dados real de processos judiciais de perdas e danos envolvendo bancos, construir um diagrama de influência que possa ser usado para
    a) predição de resultados dos processos; e
    b) indicação de decisão ótima do autor ao entrar com ação nos Juizados Especiais Cíveis ou Justiça Comum, com ou sem advogado.
3. Tornar a análise reprodutível, para que possa ser aplicada futuramente em estudos semelhantes.

1.3 Contribuições

Com este trabalho, acreditamos ter contribuído nos seguintes pontos:

1. Disseminar e incentivar a pesquisa em jurimetria como disciplina do direito.

2. Recuperar e demonstrar a importância da utilização de diagramas de influência na tentativa de solucionar problemas práticos.

3. Introdução de uma metodologia geral para elaboração de estudos em jurimetria.

4. Permitir que pesquisadores interessados no tema possam interagir, recriar e modificar as análises desenvolvidas, de acordo com seus interesses.

1.4 Organização do trabalho

A dissertação está organizada em três capítulos além da introdução.

No Capítulo 2 apresentamos a Jurimetria num contexto amplo, desenvolvemos os conceitos básicos para a construção de diagramas de influências e definimos o escopo da aplicação realizada sobre a base de dados do Tribunal de Justiça de São Paulo. No Capítulo 3, discutimos a obtenção dos dados, comentando sobre as dificuldades em baixar os processos e construir a base de dados final. O Capítulo 4 é dedicado às aplicações, e nele apresentamos a construção, ajuste e validação dos diagramas de influência, mostrando e interpretando os resultados em relação aos objetivos definidos na Seção 1.2. Finalmente, no Capítulo 5 adicionamos algumas considerações finais e sugerimos novas pesquisas que poderiam ser realizadas a partir dos resultados obtidos neste trabalho.

O documento contém também dois apêndice. No Apêndice A apresentamos conceitos básicos de grafos e redes Bayesianas. No Apêndice B, apresentamos uma descrição dos pacotes estatísticos desenvolvidos para extração, tratamento e análise dos dados.

Capítulo 2

Conceitos

> "O diagrama de influência é a radiografia do problema."
>
> Carlos Alberto de Bragança Pereira

A melhor forma de introduzir um assunto depende do conceito que se quer passar e das habilidades de oratória e cognição dos indivíduos que fazem parte da comunicação. Na matemática e na estatística, por exemplo, muitas vezes é mais útil enunciar definições, pois estas trazem uma bagagem de generalidade. Na computação, é comum a definição de certos conceitos por gênero e diferença, e.g. uma nova linguagem de programação que é orientada a objetos, mas que roda mais rápido do que a linguagem C. No direito, usualmente os conceitos são passados através de exemplos e situações-problema.

No caso da jurimetria, acreditamos que não há melhor forma de introduzir um assunto do que por meio de um exemplo. Abordaremos, no entanto, os três tipos de definições supracitados, a fim de proporcionar maior clareza ao leitor interessado. É importante que fique claro que nenhuma das afirmações colocadas são definitivas. A jurimetria é uma disciplina pouco desenvolvida e ainda não é o momento de ditar seus limites.

O capítulo está organizado em três seções. Na Seção 2.1, mostramos uma situação real e explicamos algumas questões legais em que o problema está inserido. Na Seção 2.2, mostraremos alguns conceitos básicos de jurimetria, que darão suporte à linha de pensamento abordada na aplicação. Finalmente, na Seção 2.3, discutimos sobre os diagramas de influência.

2.1 Um estudo jurimétrico

O trabalho concentra-se no estudo de uma base de dados de processos cíveis do Tribunal de Justiça de São Paulo (TJSP). Nosso objetivo é obter automaticamente informações sobre os processos e, com isso,

construir um modelo capaz de predizer os resultados, considerando diferentes situações dos processos, bem como sugerir estratégias que levariam a resultados mais satisfatórios.

No Brasil, temos três poderes: Executivo, Legislativo e Judiciário, sendo nosso foco o terceiro. O Poder Judiciário é dividido em tipos de órgãos, como os Tribunais Regionais do Trabalho (TRTs), os Tribunais Superiores (STJ e STF) e o Conselho Nacional de Justiça (CNJ). O TJSP faz parte dos Tribunais Estaduais, do Distrito Federal e dos Territórios, e é responsável por julgar, em primeira e segunda instâncias,[1] os processos cíveis, criminais, família, tributos municipais e estaduais, entre outros.

O TJSP é considerado o maior tribunal do mundo, tendo um orçamento de 8,3 bilhões de reais e 43.291 cargos de servidores, dos quais 3.383 são de magistrados.[2] Em 2014 havia aproximadamente 20,4 milhões de processos ativos aguardando julgamento.

O TJSP é também o tribunal no Brasil que mais fornece oportunidades para realização de estudos, tamanha é sua relevância social e complexidade. Pesquisas baseadas no TJSP poderiam tratar de dezenas de temas diversos: do direito civil ao direito criminal; da administração pública até a legislação; do acesso à justiça até estratégias de advogados.

Escolhemos trabalhar com um tipo de caso especial (mas não raro) do direito civil, que são os procedimentos ordinários cíveis e procedimentos nos JECs, envolvendo algumas empresas específicas como réu. As empresas escolhidas foram quatro bancos (Banco do Brasil, Bradesco, Itaú e Santander), empresas de telefonia (Claro, Nextel, Tim e Vivo), a Net e a Eletropaulo.

As ações cíveis contra essas empresas geralmente envolvem conflitos de colocação indevida de nome em cadastros de inadimplentes (e.g. Serasa), cobranças indevidas, clonagem de cartões, entre outros problemas do cotidiano. Esses casos são tão comuns que, em apenas um ano, são julgados mais de cem mil ações desse tipo, o que contribui de forma significativa para o congestionamento dos tribunais. Por esse motivo, esses processos são relevantes no contexto de administração de justiça.

As ações cíveis escolhidas podem tramitar tanto na Justica Comum quanto nos JECs. Os JECs foram criados para ajudar a conter o volume de processos, considerando que grande parte deles são conflitos repetitivos. A prerrogativa do JEC é ser mais simples e célere, oferecendo ao cidadão maior acesso à justiça, sem necessidade de gastos com custas processuais e sucumbência.[3]

[1] De forma simplificada, temos a seguinte regra: um processo judicial é analisado em primeira instância, por um *juiz* em uma *vara* específica (e.g. 40ª vara cível do Foro Central Cível da Comarca de São Paulo. O juiz dessa vara profere uma *sentença*, apresentando sua decisão sobre o conflito. Se uma ou ambas as partes ficam insatisfeitas com a decisão, elas podem recorrer, e passam dessa forma para a segunda instância, que abrange todo o estado. O processo agora é apreciado por um conjunto de magistrados, que proferem um *acórdão*, que é a decisão em segunda instância.

[2] Relatório Justiça em Números, divulgado pelo Conselho Nacional de Justiça.

[3] Custas processuais são todas as despesas dos processos judiciais. Já honorários de advogados são valores que o advogado de uma das partes recebe quando essa parte ganha um processo. Sucumbência é a exigência de que a parte que perdeu a ação num processo judicial pague as custas e honorários de advogados da parte que ganhou. A sucumbência geralmente é adequada no mundo do direito pois influencia as pessoas e empresas a entrarem com ações com pedidos que façam sentido e ajuda nos altíssimos custos dos tribunais. Além disso, os valores de honorários ajudam a remunerar

Contudo, para ajuizar uma ação nos JECs sem advogado, o valor da causa não pode superar 20 salários mínimos. Este limite é extendido a 40 salários mínimos na presença de um advogado. Na Justiça Comum (leia-se, Justiça Comum igual a não JEC), o autor é obrigado a ser acompanhado de advogado.

Para indicar o tema, foi necessário discutir a estrutura dos tribunais e a matéria dos casos analisados. Geralmente é isso que ocorre em estudos jurimétricos. Como consequência, o estatístico tem a oportunidade de aprender sobre o direito e o profissional do direito já consegue entender de forma concreta o escopo do estudo, ainda tendo acesso a alguns números interessantes que podem ser contraintuitivos. Os profissionais, agora como jurimetristas, podem elaborar soluções que extrapolam o conhecimento agregado de ambos. Essa é a beleza da jurimetria.

2.2 Jurimetria

2.2.1 Definição

A relação entre direito e estatística é antiga. Jacob Bernoulli, um dos matemáticos mais conhecidos da família, é autor da obra *Arte da Conjectura*, publicada em 1713, em que descreve diversos conceitos e resultados da teoria das probabilidades e da combinatória (Bernoulli, 1713). Seu sobrinho, Nicholas Bernoulli, defendeu uma tese de doutorado em direito, mostrando possíveis aplicações da obra do tio no universo jurídico.

A utilização da probabilidade no direito passou a ser mais conveniente após o advento do *realismo jurídico*. A corrente doutrinária majoritariamente estadunidense, que teve como principais precursores Oliver Wendel Holmes Jr., Roscoe Pound e Benjamin Cardozo, permitiu que o direito pudesse ser construído a partir das decisões dos juízes e não de um conjunto normativo formal e abstrato. A primeira obra sobre o tema é o livro *The Path of Law* (Holmes, 1897).

Partindo da ideia do realismo jurídico, o termo *jurimetrics* finalmente aparece em 1949 em (Loevinger, 1949). Loevinger, no entanto, não se preocupou em definir o termo, defendendo apenas que o direito deveria ter uma disciplina científica, rejeitando qualquer forma de teorização.

A definição atual de jurimetria foi apresentada no Capítulo 1 e é a **aplicação da estatística no estudo do direito**. Como consequência da definição, a jurimetria é uma disciplina do direito que utiliza o ferramental estatístico na tentativa de elucidar fenômenos jurídicos.

Para diferenciar a abordagem que chamaremos de clássica e a abordagem jurimétrica no estudo dos fenômenos jurídicos, vamos discutir dois princípios fundamentais que divergem nas duas formas de estudo. Outras diferenças podem ser tiradas como corolários diretos desses pontos.

parte do trabalho realizado pelo advogado. O Artigo 54. da Lei 9.099/1995 decide que nos JECs não há custas na primeira instância.

Concretude. Estudar de forma concreta significa situar o objeto de estudo no tempo e no espaço. Por exemplo, na forma clássica, estudamos o tema "direito civil" com base na norma, princípios e interpretações. Na forma jurimétrica, estudamos processos que envolvem temas específicos do direito civil, distribuídas ou julgadas em dado intervalo de tempo e em determinadas varas ou tribunais.

Uma vantagem desse tipo de abordagem é que há clareza na construção do escopo das pesquisas, o que pode levar a conclusões mais diretas. Para tornar os estudos factíveis, no entanto, também é necessário definir claramente o escopo em que as conclusões são válidas.

Utilizar a abordagem concreta não significa que não podemos pensar de forma abstrata. Para construção de modelos capazes de mensurar certas quantidade de forma adequada, ou para elaborar soluções para um problema na lei, é necessário desconstruir o fenômeno de maneira ampla. A concretude ajuda ao direcionar as abstrações e ligá-las a um objetivo pragmático.

Estocasticidade. A jurimetria assume como ponto de partida a possibilidade modelar certos fenômenos do direito como eventos aleatórios. Essa abordagem pode contrastar com a abordagem clássica, que usa o determinismo fatalista como realidade do direito, onde temos de escolher entre afirmar ou negar ou, ainda, na presença de cenários mais incertos, desistir de defender qualquer coisa pela "impossibilidade de determinar".

As vantagens em assumir aleatoriedade em modelos jurimétricos assemelham-se às vantagens da utilização da metodologia estatística em geral. Por exemplo, ao aceitar que há erros nos modelos e que não somos em geral capazes de dar respostas definitivas às nossas perguntas, abrimos um leque imenso de possibilidades, em que associamos possíveis respostas a suas respectivas verossimilhanças. Assumir que os eventos são aleatórios não significa assumir que são caóticos ou irracionais. O grande trunfo da estatística está em prover ferramentas poderosas para compreensão e controle da incerteza. A modelagem estocástica também pode ser considerada essencialmente uma generalização da modelagem determinística.

2.2.2 Áreas e tópicos

No atual estado da arte, seria complicado – ou até presunçoso – definir uma forma fechada para todas as aplicações e áreas que fazem ou não parte do território jurimétrico. Ainda assim, é possível elaborar alguns rascunhos baseados no que já foi feito até o momento, que podem servir como base para guiar novos interessados e motivar discussões de pesquisadores mais experientes.

Uma abordagem para delimitar áreas da jurimetria é a partir dos diferentes interesses dos pesquisadores. Zabala e Silveira (2014) denominam esses pontos como *os três prismas da jurimetria*. Segundo

os autores, os prismas seriam a i) elaboração legislativa e gestão pública, ii) a decisão judicial e iii) a instrução probatória.

Outra forma semelhante de definir áreas da jurimetria seria de forma escalar, nos diferentes contextos que aparecem em cada estudo. Ao invés de prismas, poderíamos pensar em *esferas*.

O processo como objeto. Na primeira esfera, teríamos o processo judicial como objeto de estudo. Nele, estaríamos interessados em dois aspectos principais, i) a prova estatística e ii) a mensuração de valores.

No primeiro ponto, o interesse é verificar se um evento de fato ocorreu (e.g. um ato ilícito), o que pode ser objeto de testes de hipóteses, construído com base em argumentos jurídicos e testado a partir da observação de conjuntos de dados. Exemplos desse tema são processos de investigação de paternidade, estudos para avaliar se algum indivíduo cometeu certo tipo de crime, entre outros (Ver DeGroot *et al.* (1986) para detalhes).

Já no segundo ponto, o interesse é a mensuração de um dano. Esse tipo de problema ocorre quando um valor de indenização não pode ser calculado de forma puramente contábil, por envolver incerteza ou subjetividade. Alguns exemplos sobre esse tema são processos que envolvem lucros cessantes, perda de uma chance e danos morais.[4] Recomendamos a leitura de Stern e Kadane (2014) como um exemplo de estudo nessa esfera.

A instituição como objeto. Na segunda esfera, temos um conjunto de processos como objeto de estudo. Tais processos podem fazer parte de um ou mais temas específicos, ou então de uma vara ou tribunal. A principal diferença em relação à primeira esfera é que nesse caso olhamos para as características de um conjunto de casos, buscando padrões, e não especificidades individuais.

Nessa esfera poderíamos utilizar as informações dos processos para inferir sobre o comportamento de um tribunal, de uma empresa ou de um tema jurídico. Além disso, seria possível elaborar modelos preditivos para antecipar eventos dos processos, como a decisão do juiz, com o intuito de desenvolver estratégias advocatícias ou auxiliar na gestão pública.

A presente pesquisa é um exemplo de análise na segunda esfera. Outro exemplo de pesquisa na segunda esfera é um estudo realizado no Centro de Estatística Aplicada do IME-USP (Bonassi *et al.* , 2006), cujo objetivo era construir uma ferramenta que auxiliasse o cliente na tomada de decisão de entrar na Justiça, num cenário de incertezas a respeito de acontecimentos futuros. Uma diferença entre as duas pesquisas é que a primeira trata de decisões a serem tomadas a partir da decisão de litigar e a segunda trata justamente da decisão de litigar.

[4] O segundo ponto levantado pode envolver o problema do *contrafactual*, que geralmente é complexo de se avaliar, pois está relacionado com perguntas sobre um evento do passado que não ocorreu. Estudos para avaliação de efeitos contrafactuais muitas vezes procuram responder perguntas como "O que teria ocorrido com A caso B não tivesse ocorrido, *dado que* na realidade observamos B e A". Ver Pearl (2009).

A sociedade como objeto. Na esfera mais ampla, temos uma sociedade ou um conjunto de sociedades como objeto. Com efeito, nos estudos da terceira esfera, não estamos interessados somente nos aspectos jurídicos de um tema, mas no funcionamento da sociedade como um todo, levando em consideração a economia, aspectos psicológicos e sociais.

Pesquisas que colocam a sociedade como objeto geralmente aplicam seus estudos para elaboração legislativa, avaliação de impacto regulatório e administração dos tribunais. Usualmente, as análises levam em conta, mas não se limitam às análises dos litígios, considerando também bases de dados relacionadas a aspectos econômicos, demográficos de saúde, etc.

Um exemplo de pesquisa nesse sentido é um projeto desenvolvido pela Associação Brasileira de Jurimetria sobre impacto do tempo dos processos relacionados à adoção na probabilidade de adoção de crianças[5]. No estudo, foram analisados processos relacionados à adoção e também o Cadastro Nacional de Adoção, mostrando que existem indícios de que o tempo elevado de tramitação de processos de destituição do poder familiar poderiam influenciar negativamente na adotabilidade da criança.

Note que todas as áreas são intimamente conectadas e, em muitos casos, seria difícil classificar um estudo a uma esfera ou outra.

2.3 Diagramas de influência

O termo "diagrama de influência" foi indroduzido em detalhe pela primeira vez em Howard (1984). Após isso, Shachter (1986), Barlow (1987) e Barlow e de Bragança Pereira (1990) desenvolveram trabalhos complementares, formalizando alguns conceitos e apresentando algoritmos. Desde sua criação, os diagramas de influência abordaram tanto a relação entre grafos e probabilidades, quanto a possibilidade de trabalhar com decisões de uma forma mais completa, se comparada ao que era possível fazer com árvores de decisão.[6]

Quase no mesmo período, Pearl (1986) trouxe a ideia dos diagramas de influência sob uma perspectiva mais computacional, com diferentes princípios e um novo nome: "redes Bayesianas" ou "redes de crença" (*belief networks*). Desde então, a comunidade acadêmica, especialmente a parte da computação, tendeu a utilizar mais o termo "redes Bayesianas" ou "redes de crença" do que "diagramas de influência"[7], apesar da ideia de unir os conceitos de grafos e distribuições de probabilidades ser similar.

É interessante notar, como veremos em seguida, que os diagramas de influência estendem os conceitos desenvolvidos na teoria de redes Bayesianas, especialmente no tratamento de problemas de decisão dentro

[5] Disponível neste link.
[6] Ver Shachter (1986) para detalhes.
[7] Uma rápida busca no Google Scholar retorna 13.900 resultados para o termo "influence diagrams" e 131.000 resultados para o termo "Bayesian networks". Consulta realizada no dia 09/06/2015.

do contexto de modelos gráficos complexos. O fato dos artigos sobre diagramas de influência terem se tornado menos populares pode ter tornado o ramo das decisões em modelos gráficos subutilizada.

Trabalhos mais recentes como Dawid (2002) e Pearl (2009) levam essa discussão para um nível maior de complexidade, tratando do problema da causalidade. É possível encontrar até mesmo artigos que trabalham com diagramas de influência causais, por exemplo em Xiaoxuan *et al.* (2013).

Nesse texto vamos utilizar a notação dada em Hu *et al.* (2012), que é bastante simplificada, mas suficiente para atender às nossas necessidades. No artigo, um diagrama de influência é uma rede Bayesiana que contém decisões e utilidade. Para detalhes e definições técnicas, ver o Apêndice A.

Capítulo 3

Dados

> "He who would search for pearls must dive below."
>
> John Dryden

Obter dados da internet é uma tarefa complicada. A maioria dos sistemas que acumulam dados na web foram concebidos para utilização gerencial ou busca de informações, não para análise de um conjunto microdados. Por essa razão, é comum a necessidade de construção de ferramentas computacionais, que podem se tornar obsoletas com simples atualizações dos sistemas.

No mundo do direito isso não é diferente. Muitos Tribunais de Justiça disponibilizam ferramentas de pesquisa que dependem de números identificadores dos processos e fazem uso de *captchas*.[1] Quando apresentam ferramentas para consulta de jurisprudência, as bases são duvidosas, contemplando apenas parte dos processos disponíveis publicamente.

Apesar das dificuldades, obter dados da web pode ser uma atividade gratificante. Como dizia Arthur Nielsen, *"The price of light is less than the cost of darkness."*. Entender de que forma os dados são organizados e armazenados nos tribunais revela muito sobre a forma de pensar do profissional do direito, e sobre o próprio direito em si. No decorrer deste trabalho, aprendemos diversos detalhes essenciais para o entendimento dos tribunais que dificilmente seriam discutidos em cursos tradicionais do direito, como, por exemplo, a disparidade nas decisões dos juízes, os maiores litigantes, os tipos de litígio com maior volume processual, entre outros.

Outra grande vantagem da extração de dados na web é que, uma vez realizada, pode ser replicada e reproduzida com um esforço menor do que o trabalho inicial. Como consequência, mais pesquisas podem aparecer, explorando diferentes aspectos do direito e gerando cada vez mais conhecimento so-

[1] "Completely Automated Public Turing test to tell Computers and Humans Apart" (teste de Turing público completamente automatizado para diferenciação entre computadores e humanos), Von Ahn *et al.* (2003). Geralmente são imagens com textos distorcidos em que o usuário precisa digitar o conteúdo para validar o preenchimento de um formulário.

bre o funcionamento do direito. Nós vivemos hoje na era do "pré-sal sociológico"[2] e o grande desafio colocado para pesquisadores do direito é o de conhecer o próprio direito através dos dados disponíveis publicamente. Isso permitiria elaborar estratégias efetivas para aprimoramento da prestação jurisdicional e da administração da justiça.

Organizamos o capítulo em três seções. Primeiro, discutimos um pouco sobre as estratégias existentes para obtenção de dados de processos judiciais, com enfoque no TJSP. Em seguida, mostramos como nossos dados foram obtidos, contemplando a obtenção dos documentos, limpeza e consolidação da base de dados final. Por fim, apresentamos uma breve análise descritiva dos dados obtidos.

3.1 Estratégias para coleta de dados nos tribunais

Nosso objetivo é extrair dados de ações de primeira instância nos JECs e Justiça Comum, num dado período, envolvendo como ré alguma das empresas entre Banco do Brasil, Bradesco, Itaú, Santander, Claro, Nextel, Tim, Vivo, Net e Eletropaulo.[3] Como é difícil obter tais processos de antemão, a estratégia adotada foi baixar todos os dados do período e, em seguida, selecionar os processos que fazem parte do escopo.

Para a realização de uma coleta estruturada e eficiente, é inevitável a adoção de estratégias. Os dados dos tribunais são muito volumosos e pouco documentados, logo realizar extrações sem planejamento pode tornar o trabalho demorado. As estratégias discutidas em seguida são o resultado de inúmeras frustrações, com alguns resultados positivos.

Em todas as estratégias apresentadas, nosso objetivo é sempre o mesmo: obter números identificadores de processos. [4] Cada estratégia apresenta vantagens e desvantagens em relação ao tempo de execução e completude da base. Os números obtidos são utilizados para consulta no sistema do TJSP, de onde finalmente conseguimos nossos dados.

3.1.1 Jurisprudência

A obtenção de informações de processos via jurisprudência consiste basicamente em utilizar as ferramentas de busca disponíveis nos tribunais. Geralmente, podemos pesquisar por palavras-chave, deter-

[2] O pré-sal brasileiro é uma área de reservas de petróleo profunda, situada abaixo de camadas de rocha salina ((Lima , 2008)). O pré-sal representa um considerável desafio tecnológico mas também uma grande oportunidade de negócio. O termo pré-sal sociológio foi criado analogamente a essa ideia, pois temos diversas bases de dados públicas e potencialmente valiosas, mas que ainda precisam de um trabalho pesado de extração e transformação de dados para serem adequadamente analisados.

[3] A escolha das empresas teve como base o volume processual observado e a popularidade das empresas, além do interesse em estudar bancos, empresas de telefonia e fornecedores de serviços essenciais, como a Net e a Eletropaulo.

[4] É suficiente no nosso caso encontrar os números de processos, pois o TJSP dispõe de ferramentas para busca desses processos, o que possibilita a captura dos dados de nosso interesse.

minadas comarcas ou varas, etc., obtendo a lista de processos que combinam com esse filtro. O TJSP é um dos únicos tribunais que permite a pesquisa dos julgados na primeira instância. A ferramenta faz parte do sistema e-SAJ do TJSP, denominada Consulta de Processos do Primeiro Grau (CJPG).

Por conta do elevado volume processual, e com o intuito de permitir a reprodução de pesquisas, construímos algumas ferramentas computacionais para extração sistemática dos documentos. Essas ferramentas fazem parte de um pacote do R dedicado exclusivamente à raspagem dos documentos de diversas fontes de dados disponíveis no TJSP. O código do pacote `tjsp` está aberto ao público e é um dos subprodutos do presente projeto de mestrado.

O programa para extração dos dados funciona como uma espécie de robô que imita as ações de um ser humano. As páginas acessadas, formulários preenchidos e cliques realizados em um navegador podem ser transcritos em requisições web, que são aplicadas diretamente por funções disponíveis no pacote, que recebem documentos como resultado, usualmente páginas HTML.[5] Os documentos obtidos são armazenados localmente para serem processados em seguida.

No caso da CJPG, o robô realiza duas tarefas básicas: busca e paginação. Para a busca, o programa precisa reproduzir o resultado de uma consulta realizada a partir do preenchimento do formulário de pesquisa. No nosso caso, para a paginação, foi necessário acessar aproximadamente 15 mil páginas de resultados (cada página de resultados na CJPG apresenta 10 julgados), mantendo a sessão de acesso utilizada na pesquisa inicial.

Após o armazenamento das páginas HTML, passamos para uma fase de raspagem dos documentos. A tarefa envolve basicamente a transformação dos dados em formato semi-estruturado (linguagem de marcação) em dados estruturados (em forma de tabela). Para isso, é necessário identificar uma forma geral para extrair cada atributo do documento e rodar o mesmo algoritmo para todos os documentos. Usualmente, essa identificação é feita a partir da aplicação de *XPaths* (*XML Path Language*), que são similares a expressões regulares.

Pesquisadores interessados em realizar estudos com diferentes especificações poderiam utilizar a função `cjpg` do pacote `tjsp` para a extração dos dados. Um breve tutorial para utilização do pacote encontra-se no Apêndice B.

Consistência da CJPG

A partir de verificações com o tribunal, fomos informados que existe uma limitação da base disponível na CJPG. Essa limitação, no entanto, não é devidamente documentada, de forma que seria complexo ou até impossível conhecer o mecanismo que identifica os processos que aparecem e que não aparecem

[5] *HyperText Markup Language*.

no sistema. Um exemplo que observamos na base é o número reduzido de processos no Foro Central dos Juizados Especiais Cíveis e a completa ausência de processos das duas Varas dos JECs no Foro Regional I de Santana.[6]

As análises realizadas sobre a base de dados obtida a partir da CJPG ignoram completamente o possível viés dos casos encontrados. Por conta disso, conclusões sobre resultados obtidos dessa forma estariam condicionadas a esta limitação.

3.1.2 Diários Oficiais

A forma mais tradicional de obtenção de informação dos tribunais é através dos Diários Oficiais. Tais Diários são disponibilizados oficialmente por todos os tribunais brasileiros e contemplam todas as tramitações processuais. Logo, todos os números de processos seriam acessíveis nos Diários Oficiais.

O Diário de Justiça Eletrônico (DJE) do TJSP é dispobilizado desde 2007 e contém diversas informações de forma não estruturada em arquivos PDF. Para obter os números de processos desses arquivos, a forma usual é trascrevê-los para XML ou texto e posteriormente aplicar máscaras de números de processos.

O pacote `tjsp` também apresenta meios para obtenção de números de processos via Diários Oficiais. No nosso estudo, realizamos o download de todos os cadernos dos Diários Oficiais no ano de 2014. O download dos Diários é uma tarefa relativamente rápida, sendo possível baixar um ano de Diários em apenas algumas horas.

Consistência dos Diários

Uma dificuldade em relação a essa abordagem é que nem todos os processos aparecem nos Diários Oficiais. Pelo que observamos, quando um processo tramita sem advogados das partes (o que pode ocorrer somente se o processo tramitar no JEC com valor da causa abaixo de vinte salários mínimos) e é solucionado através de acordos entre as partes, o processo não aparece nos Diários Oficiais. Com efeito, se o interesse do pesquisador é investigar justamente processos nos JECs (como é o caso de nosso projeto de mestrado), seria necessário considerar esse ponto[7].

Uma outra característica relacionada à obtenção dos processos por Diários Oficiais é que a pesquisa pode resultar em processos que ainda não foram julgados. Isso ocorre pois nesse caso o estudo é prospectivo, em que os processos são indexados pelas datas de início, não de julgamento. Diferentemente da CJPG, poderíamos obter processos com informações incompletas. Isso é um fator relevante caso seja de

[6] Uma lista completa de todas as varas do TJSP pode ser obtida neste link.

[7] Observe, por exemplo, o processo "0004695-06.2014.8.26.0002", que existe e é pesquisável, mas não aparece nos Diários Oficiais.

interesse da pesquisa analisar o tempo dos processos.

3.1.3 Amostragem

Podemos pensar também numa solução estatística para obtenção dos processos. O número identificador de um processo judicial utilizado atualmente pelos tribunais é chamado **número CNJ**, criado na Resolução 65 do CNJ[8]. A resolução define o padrão `NNNNNNN-DD.AAAA.J.TR.0000`, descrito abaixo.

- `NNNNNNN`: Número identificador do processo.
- `DD`: Dígito verificador, gerado a partir da aplicação do algoritmo Módulo 97 Base 10, conforme Norma ISO 7064:2003.
- `AAAA`: Ano do ajuizamento do processo.
- `J`: Segmento do poder judiciário. No nosso caso, esse número é sempre `8`, que identifica a Justiça Estadual.
- `TR`: Identifica o tribunal. No nosso caso, esse número é sempre `26`, que corresponde ao TJSP.
- `0000`: Identifica a unidade de origem do processo. No nosso caso, as possíveis configurações correspondem aos foros da comarca de São Paulo[9].

Uma utilidade interessante que a especificação do número CNJ traz é a possibilidade de gerar todos os possíveis números de processos. Para cada configuração de ano, justiça, tribunal e órgão de origem, seriam dez milhões de números distintos. No nosso caso, por exemplo, considerando somente processos ajuizados em 2014, o Foro Central Cível, Foro Central do Juizado Especial Cível e os doze principais Foros Regionais, seriam aproximadamente 140 milhões de números distintos. Esses números poderiam ser utilizados para pesquisar diretamente todos os processos do TJSP.

No entanto, pesquisar 140 milhões de números seria, na prática, uma tarefa demorada e onerosa para os tribunais. Se realizássemos, por exemplo, 10 pesquisas por segundo, seriam necessários aproximadamente 162 dias – metade de um ano – para a realização de todas as consultas. É nessa situação que uma alternativa via amostragem parece viável. Poderíamos, ao invés de pesquisar todos os processos, gerar uma amostra aleatória de números e pesquisá-los diretamente nas ferramentas de pesquisa.

O problema dessa abordagem é que nem todos os números dos processos correspondem a processos de fato. Nos tribunais, os números são gerados conforme a demanda e, portanto, apenas uma parcela dos números seria realmente manifesta.

[8] Disponível em http://www.cnj.jus.br/busca-atos-adm?documento=2748. Acesso em 01/05/2015.

[9] Na CJPG, em alguns casos, os números das unidades de origem correspondem a foros de outras comarcas. Isso ocorre quando um processo é transferido entre comarcas por questões de competência. A lista completa das unidades de origem dos tribunais estaduais encontra-se neste link

Para estudar a probabilidade de manifestação de um número de processo, realizamos uma pesquisa preliminar, realizada com 10 mil processos gerados a partir dos 140 milhões supracitados. Encontramos uma taxa de 1.6% de números de processos reais. Com efeito, para obtenção de uma amostra de tamanho aproximado n, seria necessário consultar $62.5 * n$ números gerados aleatoriamente.

Ao estudar os processos baixados a partir dos Diários Oficiais, observamos também um resultado curioso em relação aos processos existentes. Os números de processos gerados não são necessariamente sequenciais, mas apresentam alguma estrutura. Na Tabela 3.1, calculamos a proporção de processos existentes para cada combinação dos dois números iniciais dos processos. Observe que apenas algumas combinações aparecem com números expressivos. Em outras investigações, observamos que a proporção de números que começam com o dígito "1" é maior em alguns foros do que outros.

Tabela 3.1: *volume e proporções de processos de acordo com seus dois primeiros dígitos.*

Dígitos	n	%
10	370.606	65.365 %
00	160.373	28.286 %
11	28.809	5.081 %
20	7.146	1.260 %
21	17	0.003 %

A redução no número de casas variando livremente nos números de processos reduz de forma expressiva a quantidade total de números possíveis. Por exemplo, considerando somente processos começados em "00", "10" e "11", o número total de processos no nosso caso cairia de 140 milhões para 4,2 milhões de processos. Como consequência, a probabilidade de gerar um processo real aumentaria, tornando uma pesquisa por amostragem factível.

Consistência dos dados obtidos por amostragem. A principal vantagem da pesquisa por amostragem é que esta superaria qualquer meio de manipulação dos dados. Por isso, poderia ser considerada mais "confiável". Sobre este ponto, três ressalvas. Primeiro, existem processos válidos que não seguem a regra dos dígitos verificadores da numeração CNJ; estes números são, no entanto, raríssimos (menos de 0.0001%, estimado a partir dos números dos Diários Oficiais de 2014). Segundo, a utilização do "atalho" de fixar alguns números iniciais para limitação no espaço amostral poderia configurar a possibilidade de manipulação, mas pelo que observamos nos estudos preliminares, a probabilidade desse evento seria muito pequena. Terceiro, pesquisar a partir de números ignora processos que foram distribuídos fora da relação de foros, mas julgados em um foro que está dentro da relação. Ou seja, nesse caso, a população de interesse seriam os processos distribuídos no foro, não os processos julgados no foro. A terceira

ressalva também é válida para a estratégia dos Diários Oficiais.

3.1.4 Resumo

A Tabela 3.2 mostra as características de cada abordagem para obtenção de processos. Cada abordagem tem seus prós e contras. Caso o tempo não seja um problema, recomenda-se a pesquisa por amostragem ou pelo DJE. Caso seja necessário pesquisar em um grande número de foros (por exemplo, todas as comarcas do Estado de São Paulo), a pesquisa por amostragem fica infactível. Caso o pesquisador não se importe com o possível viés nos dados obtidos pela jurisprudência, esta é a abordagem mais adequada.

Tabela 3.2: *características das estratégias para obtenção de números CNJ.*

Estratégia	Tempo	Tipo de estudo	Viés
Jurisprudência	Rápido	Retrospectivo	Sim, desconhecido
DJE	Médio	Prospectivo	Sim, conhecido
Amostragem	Lento	Prospectivo	Não

3.1.5 Consulta de Processos do Primeiro Grau

Com base nos números de processos levantados, construímos um segundo robô, capaz de pesquisar um determinado processo no tribunal. No caso da estratégia por amostragem, esse é um passo obrigatório para obtenção dos números válidos. As duas tarefas (teste de existência e download dos processos) são, portanto, realizados simultaneamente. Nos outros casos, é necessário rodar o robô a partir dos números levantados.

A raspagem dos dados da Consulta de Processos de Primeiro Grau (CPO-PG) representa um desafio maior do que na CJPG. A estrutura interna dos documentos HTML não é homogênea e existem muitas formas diferentes em que as partes são classificadas nos processos. Por conta disso, a base estruturada foi armazenada em três conjuntos de dados distintos: informações básicas, partes e movimentações.

O pacote `esaj` possui ferramentas para download e raspagem dos documentos obtidos via CPO-PG. A função `cpo_pg` do pacote `tjsp` é utilizada para download, e a função `parse_cpo_pg` para raspagem. Um breve tutorial para utilização do pacote encontra-se no Apêndice B.

3.2 Nossa base de dados

Para a coleta de nossa base de dados, passamos por todas as três estratégias. Partimos da mais simples, via CJPG, depois para Diários Oficiais e então para amostragem. Os problemas de consistência das bases de dados foram identificados na prática, a partir de diversas verificações.

Apesar dos esforços em obter a base de dados pela amostra, optamos por utilizar na análise a base de dados obtida via CJPG, por apresentar maior facilidade para reprodução das análises e pelo fato de apresentar dados completos (por ser retrospectiva).[10] A partir dos documentos baixados de 150 mil processos, realizamos a raspagem para obtenção da base estruturada, mas ainda em formato inadequado para análise. Começamos então a limpeza dos dados.

A tarefa de limpeza e consolidação dos dados foi a parte mais trabalhosa e dispendiosa do projeto. O processo exigiu a leitura de diversas sentenças, aplicação experimental de algoritmos e codificação específica à base de dados inicial obtida através dos algoritmos computacionais.

Um dos complicadores para a limpeza dos dados ocorre pela própria origem. Como explicamos anteriormente, as bases do TJSP foram concebidas para fins administrativos e busca processual, e não análise estatística. Por conta disso, observamos campos sem padronização como, por exemplo, o resultado dos processos, que só seriam acessíveis a partir da leitura e interpretação dos textos das decisões.

Um segundo problema diz respeito à falta de parâmetros bem definidos para buscar variáveis importantes a partir do que estava disponível. Essa dificuldade é especialmente complexa nos textos das decisões, pois os textos variam muito no tamanho e forma de escrita e, portanto, é complicado determinar padrões estruturais. Não foi possível definir *a priori* exatamente quais variáveis buscar nos textos e de que forma elas apareceriam e, por isso, tivemos várias tentativas frustadas de captura de dados.

Em terceiro lugar, tivemos dificuldade com o complexo vocabulário que é inerente à área do direito. Frases com mesmo significado eram escritas de diversas formas, sendo necessário utilizar muitas expressões regulares para atingir um resultado aceitável.

Colocados os problemas, construímos a base de dados final em três passos: informações básicas, partes e textos/movimentações. Em relação aos textos, a abordagem que teve resultados mais frutíferos foi separar todas as frases dos documentos para depois analisar os textos. Dessa forma, foi possível encontrar, de forma aproximada, quais frases estavam relacionadas ao resultado do processo. Em seguida, aplicamos diversas expressões regulares para extrair outras informações relevantes, como tipo de dano, colocação do nome em órgãos de proteção de crédito, gratuidade judiciária, relação de consumo, entre outras.

Para obtenção dos valores de condenação utilizamos expressões regulares para extração de números

[10] O possível viés da base foi ignorado na pesquisa, e seria um interessante objeto de estudo em trabalhos futuros.

e também fizemos algumas suposições sobre os textos. Por exemplo, assumimos que os primeiros valores mencionados após o proferimento da sentença eram relacionados a indenizações por danos materiais e morais, e que os valores mencionados em seguida eram provenientes de custas processuais. Existem casos em que o valor final é negativo, representando pagamento de sucumbência, que só ocorre nos casos da Justiça Comum.

O vocabulário utilizado pelos magistrados e escreventes no proferimeto de sentenças é, ao mesmo tempo, rico e limitado. É rico por conta do emprego de diversas palavras que não são comumente utilizadas na língua portuguesa e é limitado pois os contextos em que ocorrem as sentenças são repetitivos.

Um ponto importante a levantar é que tratamos do processo somente no intervalo entre a distribuição e a primeira sentença. Isso exclui, por exemplo, acordos realizados pós sentença e, naturalmente, reformas de decisão em instâncias superiores.

Um problema relacionado a isso que foi encontrado nos dados durante a análise é que nem sempre as sentenças obtidas julgam o mérito ou homologam acordos. Em alguns casos, os documentos têm a finalidade de extinguir a execução do processo, ou seja, referem-se a um momento posterior à primeira decisão, o que fugiria do escopo do estudo. A análise dos textos também foi importante para identificação desses documentos e exclusão desses processos da base de dados final.

Reprodutibilidade

Para uma descrição mais completa dos métodos aplicados para limpeza e consolidação da base de dados final, recomendamos a leitura dos códigos disponíveis no pacote `tjsp.data`. Neste documento nos restringimos a descrever somente os pontos principais, e os códigos podem revelar outras suposições assumidas que podem ser relevantes para essa e futuras pesquisas.

No início do projeto, um dos objetivos era que a pesquisa fosse não só reprodutível, mas também replicável, ou seja, que o estudo pudesse ser realizado novamente com diferentes especificações como classes processuais, cortes temporais etc. No entando por conta da dificuldades na construção de um procedimento geral na limpeza dos dados, esse objetivo acabou sendo atingido de maneira restrita. Dessa forma, um pesquisador que tiver interesse em reproduzir a pesquisa poderá aplicar diretamente os algoritmos de download, raspagem e modelagem dos dados, mas terá de adaptar os algoritmos de limpeza para que a qualidade dos dados não fique comprometida.

3.2.1 Base de dados final

A base de dados final contém 19.078 processos e 13 variáveis, especificamente:

- `n_processo`: Número do processo.

- `cod_sentenca`: Código interno da sentença.
- `foro`: Nome do foro.
- `empresa`: Empresa envolvida (banco, telefonia ou outros).
- `tipo_vara`: Tipo de vara (vara cível, JEC com advogado, JEC sem advogado).
- `tipo_dano`: Tipo de dano (dano material, moral ou ambos).
- `resultado`: Resultado do processo (acordo, procedente, improcedente, parcialmente procedente).
- `valor_acao`: Valor da ação, em salários mínimos.
- `resultado_vl`: Valor do resultado (valor de condenação ou do acordo), em salários mínimos.
- `tempo`: Tempo do processo desde a distribuição até a sentença, em anos.
- `serasa`: Indicador de presença de discussões a respeito dos órgãos de proteção ao crédito.
- `consumo`: Indicador de presença de discussões a respeito de relação de consumo.
- `gratuidade`: Indicador de presença de discussões a respeito de gratuidade judiciária.

Além das variáveis mencionadas, temos outras informações nas pases do pacote `tjsp.data`. Essas variáveis foram omitidas neste texto por serem menos relevantes para o presente estudo.

3.3 Análise descritiva

Nessa breve análise descritiva colocamos alguns pontos principais que podem ser relevantes ou interessantes.

A Tabela 3.3 mostra a distribuição dos procesos em relação ao tipo de vara. Podemos notar que dois terços dos casos correspondem a procedimentos na justiça comum, enquanto nos JECs, temos uma proporção semelhante de processos com e sem advogado do autor. O resultado é contraintuitivo, pois esperaríamos que o número de casos nos JECs fosse maior para esse tipo de caso. Isso pode decorrer da forma de amostragem utilizada, que ignorou processos de algumas varas específicas.[11]

Tabela 3.3: *volume e proporção de processos em relação aos tipos de vara.*

tipo_vara	n	%
Comum com advogado	12427	65.1%
JEC com advogado	3595	18.8%
JEC sem advogado	3056	16.0%
Total	19078	100%

[11] Essa afirmação poderia ser testada a partir dos dados obtidos via amostragem. A verificação dessa e de outras possíveis fontes de viés serão tópicos de trabalhos futuros.

A Tabela 3.4 mostra a distribuição dos processos em relação à empresa. Note que os bancos dominam o volume processual, seguidos pelas três principais empresas de telefonia móvel. É interessante notar que o Itaú é réu em mais que o dobro de processos envolvendo o Banco do Brasil. Uma possível justificativa seria a elevada quantidade de serviços que o Itaú fornece ao consumidor, e também a fusão com o Unibanco, que acabou agrupando os processos em uma só empresa.

Tabela 3.4: *volume e proporção de processos em relação às empresas.*

empresa	n	%
ITAU	4852	25.4%
BRADESCO	3032	15.9%
SANTANDER	2443	12.8%
BB	2202	11.5%
VIVO	1882	9.9%
CLARO	1580	8.3%
TIM	1183	6.2%
ELETROPAULO	783	4.1%
NET	622	3.3%
NEXTEL	499	2.6%
Total	19078	100%

A Tabela 3.5 mostra a distribuição dos processos em relação ao tipo de dano. Aqui observamos uma grande quantidade de informações faltantes nos nossos dados.

Tabela 3.5: *volume e proporção de processos em relação aos tipos de dano.*

tipo_dano	n	%
dano moral	6647	34.8%
NA	6462	33.9%
dano moral e material	4110	21.5%
dano material	1859	9.7%
Total	19078	100%

A Tabela 3.6 mostra a distribuição dos resultados dos processos. Observamos que o resultado mais frequente é a procedência da ação. Também, ainda que o acordo seja a categoria menos frequente, sua proporção é significativa.

Tabela 3.6: *volume e proporção de processos em relação aos resultados.*

resultado	n	%
Procedente	5868	30.8%
Parcialmente	4795	25.1%
Improcedente	4549	23.8%
Acordo	3866	20.3%
Total	19078	100%

Como podemos observar na Tabela 3.7, esses resultados parecem ser associados ao tipo de vara. Note que a proporção de acordos é maior nos JECs. Nos JEC's, a presença de advogado diminui a chance de acordo e aumenta a de procedência parcial.

Tabela 3.7: *Volume e proporção de processos em relação aos resultados para cada tipo de vara.*

tipo_vara	Acordo	Improcedente	Parcialmente	Procedente	Total
Comum com advogado	12.5%	29.7%	25.6%	32.1%	100.0%
JEC com advogado	31.4%	12.4%	28.2%	28.1%	100.0%
JEC sem advogado	38.6%	13.4%	19.8%	28.2%	100.0%

A Tabela 3.8 relaciona a variável serasa ao resultado do processo. Mesmo a variável não identificando a colocação do nome da parte em órgãos de proteção ao crédito, observamos sua associação com o resultado da ação. Quando a variável assume o valor sim, a proporção de casos parcial ou totalmente favoráveis ao autor é maior que 70% enquanto que a proporção de acordos é menor que 10%. A proporção de casos em que aparece serasa é de 35,7%.

Tabela 3.8: *Volume e proporção de processos em relação aos resultados em processos em que houve/não houve colocação do nome da pessoa no Serasa.*

serasa	Acordo	Improcedente	Parcialmente	Procedente	Total
não	27.71%	25.80%	22.20%	24.29%	100.0%
sim	6.8%	20.3%	30.4%	42.4%	100.0%

Em relação aos valores das ações, é importante destacar dois pontos importantes. Primeiro, o valor da ação nos processos do JEC são limitados a 20 salários mínimos quando o autor não tem advogado, e a 40 salários mínimos quando o autor tem advogado. A Figura 3.1 mostra isso de forma mais clara. Observe que, no caso do JEC sem advogado, a distribuição é assimétrica, com valores próximos de zero ou próximos ao valor máximo permitido. Já nos casos de JEC com advogado, a distribuição parece

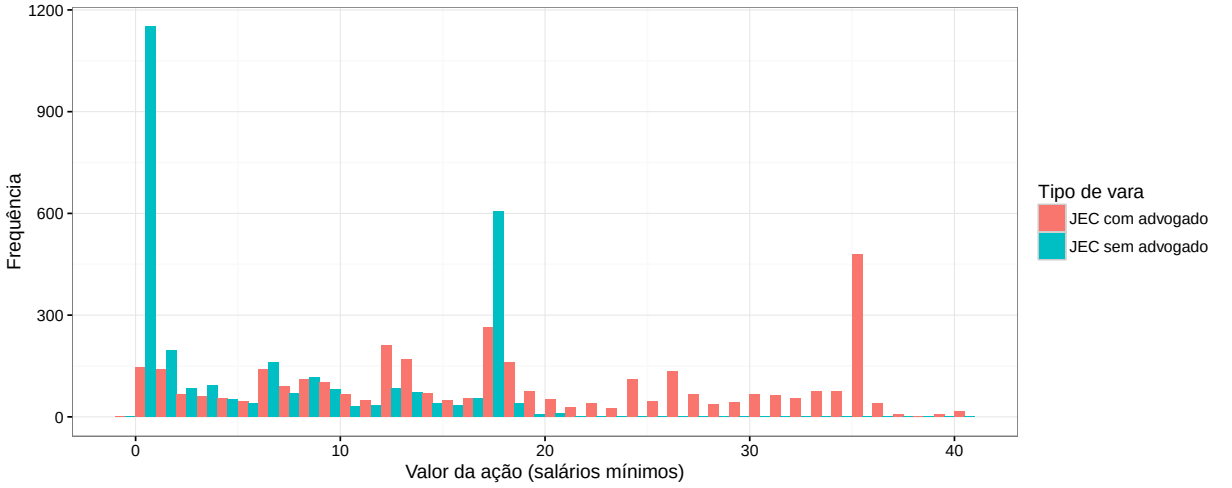

Figura 3.1: *Distribuição dos valores das ações nos JECs.*

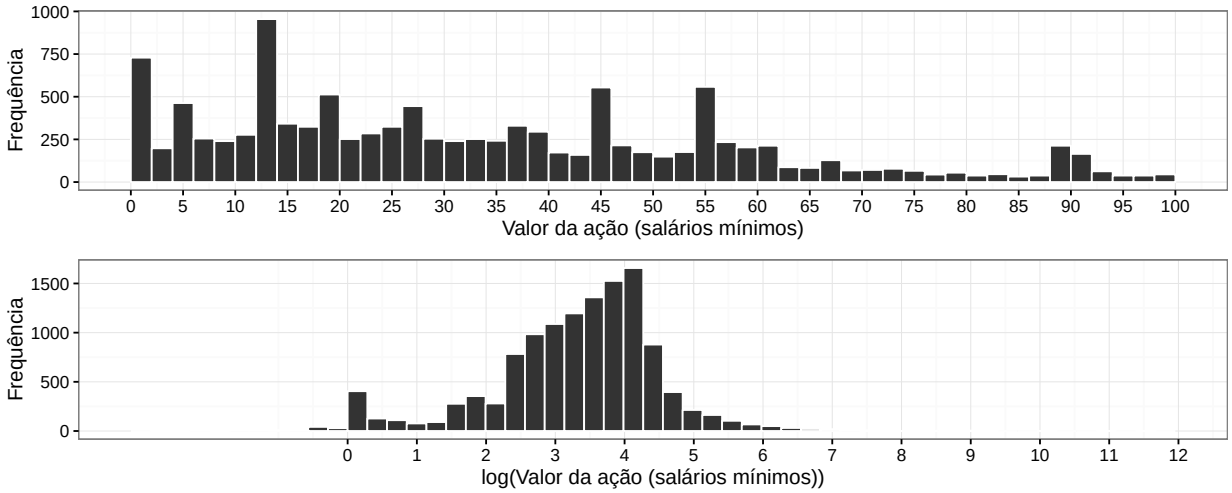

Figura 3.2: *Distribuição dos valores de ação na Justiça Comum (acima) e distribuição dos logaritmos dos valores de ação (abaixo).*

ser mais uniforme, com alguns picos em valores específicos. Observamos nos dados muitos casos de arredondamento dos valores, por exemplo, em 10 mil ou 20 mil reais.

Na Justiça Comum, a assimetria é tão grande que a visualização fica prejudicada sem uma transformação. A Figura 3.2 mostra dois gráficos. O gráfico da parte de cima é o histograma do valor da ação, limitando o gráfico para no máximo 100 salários mínimos (omitidas 1002 observações). O gráfico da parte de baixo é o histograma do logaritmo do valor da ação. É possível observar que existem concentrações por volta de zero, 12 (doze), 45 (quarenta e cinco) e 55 (cinquenta e cinco) e 90 (noventa) salários mínimos. A distribuição do logaritmo do valor da ação aproxima-se de uma normal, exceto pelo pico observado em torno de zero. O outro pico concentra-se em torno de 55 (cinquenta e cinco) salários mínimos.

Em relação aos valores de indenização, observamos, uma alta taxa de dados faltantes. São quase 40% de observações omissas. Isso ocorre por conta da dificuldade em obter esses valores de forma automática

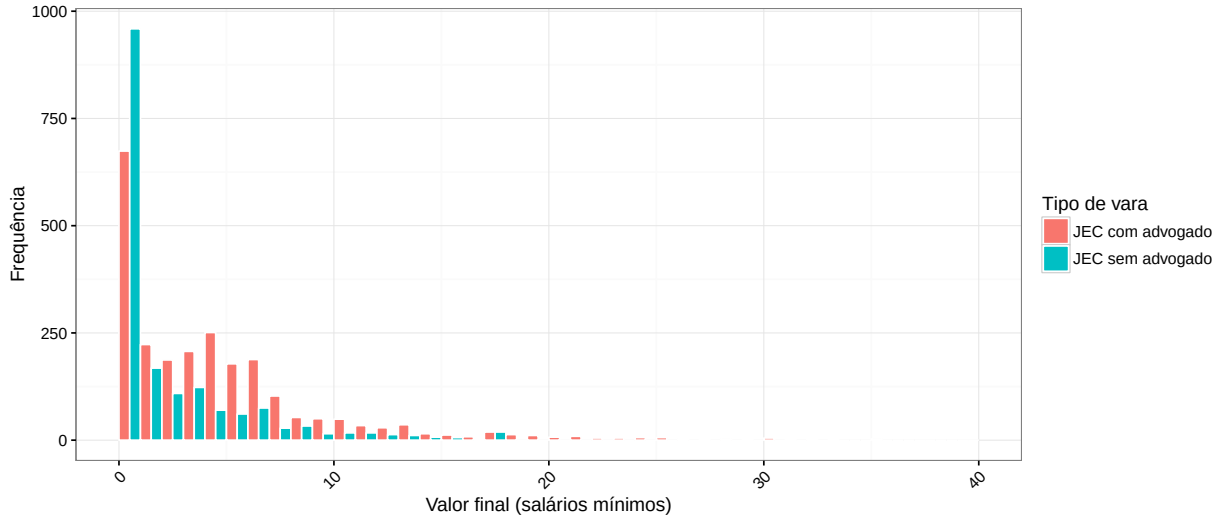

Figura 3.3: *Distribuição dos valores de indenização para processos nos JECs.*

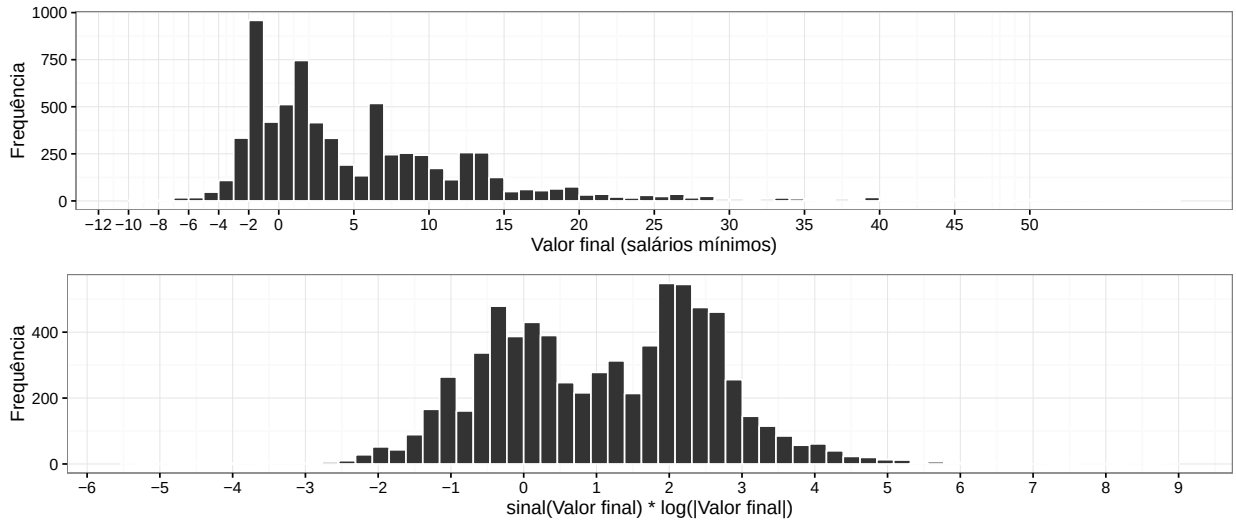

Figura 3.4: *Distribuição dos valores de indenização transformados na Justiça Comum.*

a partir das sentenças. A Figura 3.3 mostra os valores finais comparando-se JEC com advogado e JEC sem advogado. Observamos que JEC com advogado apresenta valores de indenização mais elevados do que no caso de JEC sem advogado. Observamos também poucos casos com resultados acima de dez salários mínimos.

Para visualizar os valores finais em processos da Justiça Comum, aplicamos a escala logarítmica ao módulo dos dados, e depois reaplicamos o sinal. Podemos observar na Figura 3.4 superior picos em -2 (menos dois), 1 (um) e 6 (seis) salários mínimos. No gráfico inferior, observamos uma distribuição bimodal, com concentrações em valores negativos menores que -2 (menos dois) salários mínimos e valores positivos por volta de 8 (oito) salários mínimos.

A Figura 3.5 apresenta o gráfico de dispersão das duas variáveis relacionadas ao valor. Os valores foram limitados em até cem salários mínimos. Foram omitidas 643 das observações, além das 7569

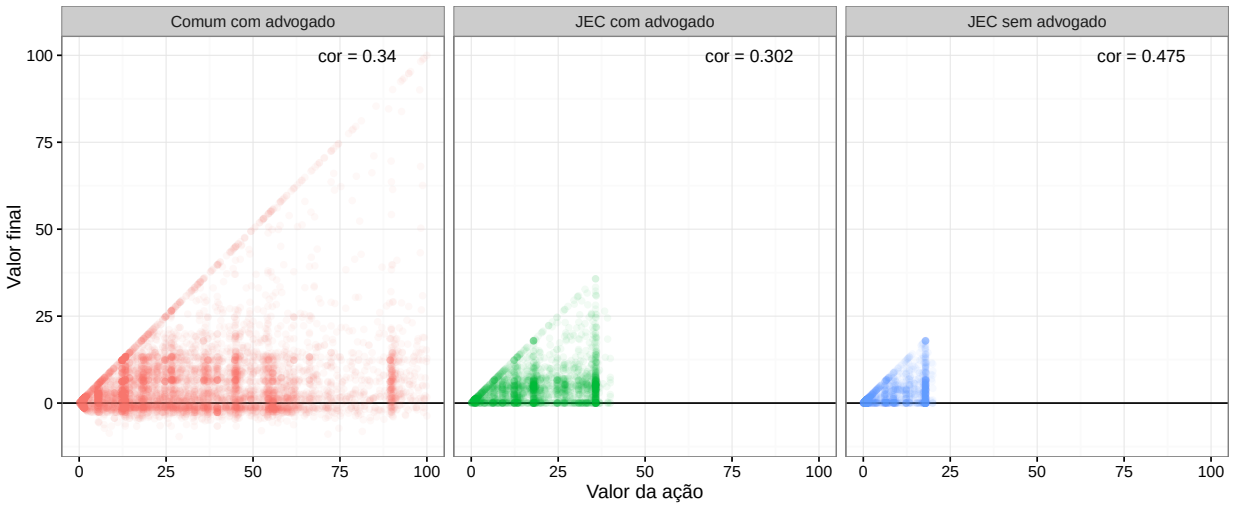

Figura 3.5: *Dispersão dos valores da ação e de indenização.*

com dados omissos para valor final. Podemos notar uma relação peculiar a limitação dos valores de indenização, que precisam ser menores que o valor da causa . Além disso, notamos que, nos JECs, o valor da ação é limitado superiormente e o valor de indenização nunca é negativo. Em todas os tipos de vara, observamos que o valor de indenização dificilmente ultrapassa 20 (vinte) salários mínimos, independentemente do valor da ação.

Capítulo 4

Aplicações

As análises que seguem podem ser entendidas como uma amostra das possíveis situações em que nosso modelo pode contribuir no âmbito jurídico. A estratégia utilizada para chegar aos resultados passa por quatro passos, que são nossas seções: especificação, ajuste, predição e decisão.

Na Seção 4.1, construímos um diagrama de influências com base nos dados obtidos e discutimos as principais suposições utilizadas para determinação das distribuições das variáveis, prioris e independências condicionais. Também mostramos outros modelos que foram utilizados para obtenção dos resultados.

Na Seção 4.2 mostramos como foi realizado o ajuste do modelo proposto. Nessa parte também explicamos as técnicas utilizadas para verificação de adequação do modelo ajustado.

Na Seção 4.3 realizamos uma série de pequenos estudos observando as predições do modelo em relação à diferentes especificações. Com base nos resultados, discutimos algumas possíveis aplicações do nosso modelo.

Na Seção 4.4 unimos os conceitos de redes Bayesianas e teoria da decisão. Com base no nosso modelo, propomos uma função de perda e tentamos responder à pergunta: "em quais casos vale mais à pena entrar no JEC e em quais casos vale mais à pena entrar na Justiça Comum?"

4.1 Especificação

Com base na análise descritiva e em alguns conhecimentos prévios sobre o tema (explicados abaixo), construímos uma rede Bayesiana contendo 10 variáveis. O modelo pode ser visualizado na Figura 4.1. As variáveis discretas são indicadas por círculos e as variáveis contínuas por hexágonos.

4.1.1 Dependências condicionais

A empresa é a variável que inicia a rede.

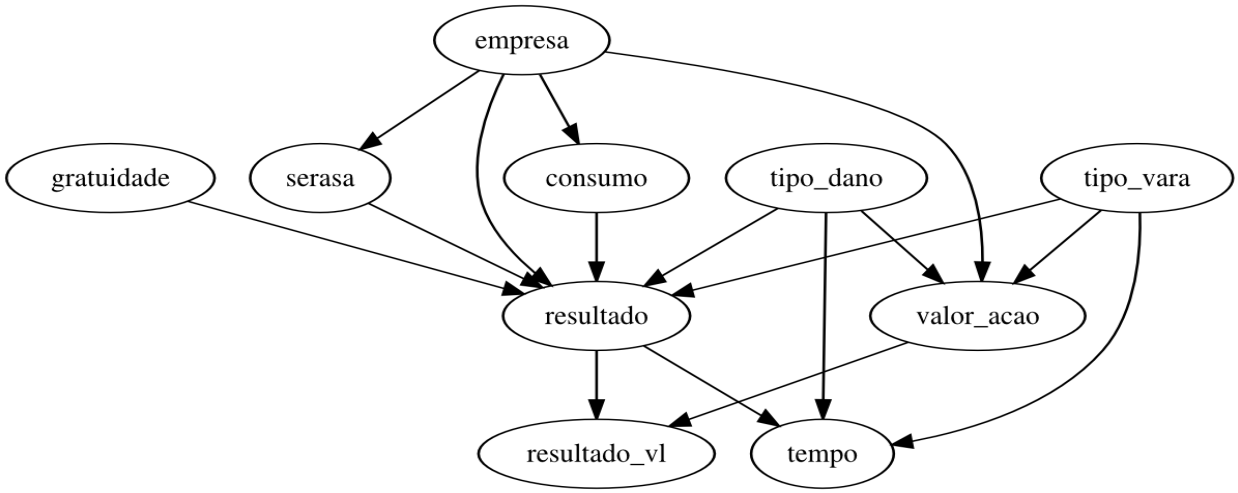

Figura 4.1: *Rede Bayesiana proposta.*

O tipo de vara afeta diretamente o resultado e o tempo. Essa variável é de interesse no estudo, pois será utilizada na parte que discutiremos sobre decisões. Como não consideramos nenhum pai para o tipo de vara, essa variável, quando considerada no conceito de decisões, pode ser interpretada como uma ação externa no modelo (Pearl, 2009).

Em relação ao valor da ação, temos dependências diretas do tipo de dano, tipo de vara e da empresa. Imaginamos que isso aconteça pois pode existir uma relação do tamanho de uma empresa com o valor pedido. Além disso, os tipos de dano tendem a delimitar o valor da ação, pois, se temos dano material, o valor é dado pelo preço de mercado e, se temos dano moral, os valores tendem a ficar próximos dos limites do tipo de vara (no caso dos JECs).

Em relação ao resultado, colocamos diversas variáveis pois, de fato, praticamente todas as informações do processo afetam seu resultado. Observe que, com exceção do valor da ação, todas as variáveis que não são consequências do resultado (tempo e valor de indenização), são pais de resultado. Não incluímos o valor da ação pois assumimos que este só tem influência sobre o valor de indenização.[1]

O tempo depende diretamente do tipo de vara, do resultado e do tipo de dano. Colocamos essas dependências pois i) processos que acabam em acordo usualmente são mais céleres; ii) o tipo de vara influenciaria o tempo, já que um dos maiores motivos para criação dos JECs era tornar o processo mais célere; e ii) o tipo de dano pode levar à necessidade de produção de diferentes provas e, portanto, a diferentes tempos.

As prioris foram determidadas da forma mais simplificada possível. Neste trabalho, não foi feita elicitação de prioris nem avaliada a sensibilidade dos resultados em relação a essas prioris. Isso configura uma limitação do estudo, e é um dos possíveis desdobramentos a serem desenvolvidos no futuro.

[1] O valor da ação poderia influenciar na decisão entre procedência total e parcial. Processos com valores muito altos estariam mais sujeitos à procedência parcial. Essa possibilidade foi ignorada no nosso modelo.

Assumimos prioris independentes para cada configuração dos pais discretos de cada variável. Por exemplo, no caso do resultado, temos $3 * 2 * 2 * 3 * 2 * 3 = 216$ distribuições Dirichlet com 4 categorias, cada uma referindo-se a uma combinação dos seus pais.

Em relação às variáveis discretas, assumimos sempre prioris Dirichlet com todos os parâmetros iguais a 1. Assim, as distribuições Dirichlet são todas uniformes nos *simplex* definidos de acordo com o número de categorias de cada variável.

Em relação às variáveis contínuas, assumimos as prioris conjugadas normal-gama-inversa. As médias dos parâmetros de regressão são iguais a zero e a matriz de covariâncias é igual a mil vezes a matriz identidade. Na literatura, existem prioris mais elaboradas para os parâmetros de regressão como, por exemplo, as prioris definidas no pacote `deal`. No entanto, escolhemos prioris mais simplificadas por conveniência.

4.1.2 Outros modelos

Para realizar comparações com nosso modelo, ajustamos um modelo de redes Bayesianas com inferência clássica e um modelo de florestas aleatórias[2].

O modelo de redes Bayesianas foi ajustado com o pacote `bnlearn`. Como o pacote não trabalha com dados omissos, utilizamos uma técnica chamada *imputação múltipla de dados*, utilizando o pacote `Amelia`[3]. Os resultados do modelo foram utilizados para que fosse possível comparar o poder preditivo e as decisões do modelo proposto com alguma outra técnica.

O objetivo do modelo de florestas aleatórias é tão somente o de comparar seu poder preditivo com relação aos resultados dos processos com o poder preditivo do modelo proposto. Para isso, utilizamos o pacote `caret` e consideramos como variável resposta o `resultado`. Como variáveis explicativas, consideramos os pais de `resultado` descritos na Figura 4.1, a saber, `gratuidade`, `tipo_vara`, `tipo_dano`, `serasa`, `empresa` e `consumo`. Para resolver o problema dos dados omissos na variável `tipo_vara`, consideramos todos os valores omissos como pertencendo a uma nova categoria, o que só foi possível pois todos os pais de `resultado` são discretos. Com relação aos parâmetros do modelo, consideramos todos os valores padrão do método de florestas aleatórias do pacote `caret`.

[2]Florestas aleatórias compõem uma classe de modelos bastante populares para construção de modelos preditivos. Ver Friedman *et al.* (2001) para detalhes.

[3]O método amelia II é um método de imputação múltipla bastante popular nas aplicações em estudos observacionais, especialmente nas Ciências Sociais. Ver Gary King et. al.

4.2 Ajuste

Nosso modelo foi ajustado utilizando-se um amostrador de Gibbs em dois passos, como descrito na Seção A.4. No primeiro passo, geramos valores para as variáveis com dados faltantes e, no segundo passo, geramos valores para os parâmetros, a partir das condicionais completas.

É importante notar que só pudemos utilizar o amostrador de Gibbs neste caso pois consideramos prioris conjugadas para nossos parâmetros. Se tivéssemos considerado uma distribuição não conjugada para os parâmetros das variáveis contínuas, ou então se considerássemos em nossa rede alguma variável contínua como pai de uma variável discreta, por exemplo, seria necessário utilizar outras metodologias para gerar amostas da distribuição à posteriori.

Nosso modelo tem, no total, 1025 parâmetros, incluindo aqueles que são determinados pela restrição da distribuição Dirichlet. A maior parte dessas quantidades se deve à variável `resultado` que, por conta do elevado número de combinações dos valores de seus pais, resulta num total de 864 parâmetros.

Também precisamos fazer algumas manipulações nas variáveis contínuas para facilitar o ajuste do modelo. Para a variável `valor_acao`, aplicamos a transformação $\log(x+1)$. Para a variável `resultado_vl`, por possuir valores negativos, aplicamos a transformação $f(x) = sign(x)\log(|x|+1)$. Na variável tempo, aplicamos a função raiz-quadrada.[4]

Para realização de testes do poder preditivo do modelo utilizamos a estratégia *k-folds* com $k = 8$. A técnica consiste em particionar a base de dados, de forma aleatória nas partições $P = \{P_1, \ldots, P_k\}$ e, para cada partição P_j, ajustamos o modelo para a base complementar $P\ P_j$, avaliando as proporções de acertos na base de dados da partição P_j. Observe que esta estratégia é uma generalização da técnica *leave one out*, que pode ser obtida se $k = n$, com n o número de observações da amostra.

Nosso amostrador de Gibbs rodou por aproximadamente três horas, gerando 10 mil amostras dos parâmetros para cada partição da base de dados. Não foi necessário aplicar muitas iterações pois o modelo, não fosse as variáveis com omissão, já estaria gerando realizações independentes da distribuição à posteriori (se não houvesse omissão nas variáveis, a inferência seria exata). Aplicamos um *burn-in* de duas mil observações e não aplicamos *thinning*.

Os parâmetros iniciais só foram necessários para geração dos primeiros valores de substituição nas variáveis com omissão. Por esse motivo, para determinar os valores iniciais da cadeia utilizamos estimativas pontuais baseadas em estatísticas da amostra, desconsiderando-se os dados omissos. Por exemplo, para definição dos valores iniciais dos parâmetros da variável `tipo_dano` calculamos sua tabela de contagens,

[4] A função raiz-quadrada é mais adequada do que a transformação logarítmica nesse caso pois é menos agressiva. Uma transformação logarítmica poderia tornar diferenças entre tempos quase irrelevantes para o modelo, principalmente na parte de decisão.

desconsiderando os valores omissos.

4.3 Predição

Para validação do modelo, estimamos seu poder preditivo em relação à variável `resultado`, considerando como informação disponível os pais de resultado. Para isso, como mencionado acima, particionamos a base em oito e obtivemos as predições para cada uma dessas partições.

As predições são feitas da seguinte maneira. Nosso objetivo é calcular a probabilidade de um novo evento $\tilde{x}$, dadas as observações $\mathbf{X} = \mathbf{x}$. Intuitivamente, podemos esperar que $\tilde{X}$ e $\mathbf{X}$ são independentes, mas na realidade essas quantidades são apenas condicionalmente independentes dado o vetor de parâmetros $\boldsymbol{\theta}$. Assim, temos

$$p(\tilde{R} = \tilde{r}|\mathbf{X} = \mathbf{x}) = \int_{\boldsymbol{\theta}} p(\tilde{X} = \tilde{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X} = \mathbf{x}) d\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}}\Big[p(\tilde{X} = \tilde{x}|\boldsymbol{\theta})\Big].$$

No nosso caso específico, gostaríamos de calcular a probabilidade do evento $\tilde{R} = \tilde{r}|\tilde{\mathbf{E}} = \tilde{\mathbf{e}}$, em que $\tilde{R}$ é o resultado e $\tilde{\mathbf{E}}$ é o vetor de *evidências*, ou seja, informações novas obtidas sobre algumas variáveis. Dessa forma, temos

$$p(\tilde{R} = \tilde{r}|\tilde{\mathbf{E}} = \tilde{\mathbf{e}}, \mathbf{X} = \mathbf{x}) = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}}\Big[p(\tilde{R} = \tilde{r}|\tilde{\mathbf{E}} = \tilde{\mathbf{e}}, \boldsymbol{\theta})\Big]$$

Como temos uma amostra da posteriori $\boldsymbol{\theta}|\mathbf{X} = \mathbf{x}$, podemos estimar essa quantidade fazendo

$$\hat{p}(\tilde{X} = \tilde{x}|\tilde{\mathbf{E}} = \tilde{\mathbf{e}}, \mathbf{X} = \mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} p(\tilde{R} = \tilde{r}|\tilde{\mathbf{E}} = \tilde{\mathbf{e}}, \boldsymbol{\theta}_i).$$

Em geral, é possível calcular $p(\tilde{R} = \tilde{r}|\tilde{\mathbf{E}} = \tilde{\mathbf{e}}, \boldsymbol{\theta}_i)$ pois, definida a rede Bayesiana, temos a distribuição conjunta de todas as variáveis e, portanto, poderíamos calcular qualquer evento que envolva essas variáveis. Convenientemente, o conjunto de evidências que vamos considerar são justamente os pais de `resultado`. Como efeito, $\tilde{R}|\tilde{\mathbf{E}} = \tilde{\mathbf{e}}, \boldsymbol{\theta}_i$ tem distribuição categórica com probabilidades $\boldsymbol{\theta}_i$.

Dessa forma, a conta fica: em cada partição, para cada observação da base de teste, calculamos a média dos $\boldsymbol{\theta}$'s obtidos da amostragem à posteriori. O valor predito para a observação é a categoria que, em média, apresenta a maior probabilidade.[5] Finalmente, comparamos os valores preditos com os resultados observados, obtendo a proporção de acertos.

[5] Esse é o resultado usando-se função de perda 0/1. É possível que exista uma função de perda mais natural. Por exemplo, predizer um resultado parcialmente procedente quando era procedente é menos ruim do que predizer um resultado improcedente, nesse caso. Testamos algumas funções de perda, mas para todas as funções consideradas o poder preditivo foi menor do que a perda 0/1.

Para obter os resultados do modelo de redes Bayesianas clássico, foi necessário apenas calcular as probabilidades com base nos parâmetros ajustados. Em relação ao modelo de florestas aleatórias, utilizamos a função de predição do pacote `caret`.

A Tabela 4.1 mostra os resultados comparando os três modelos para cada partição dos dados. Observamos que o modelo proposto tem resultados similares em relação aos outros modelos, em todas as partições da base. O modelo de redes Bayesianas clássico apresenta, em média, os melhores resultados. A tabela aponta para resultados satisfatórios, uma vez que foram baseados numa base de dados obtida de forma automática diretamente dos tribunais. Além disso, como existem 4 categorias e a categoria mais frequente ocorre em 30% dos casos, o ganho de acurácia em relação ao classificador trivial foi de cerca de 20%.

Tabela 4.1: *Proporções de acerto do modelo proposto, florestas aleatórias e rede Bayesiana clássica, para cada partição da base.*

Partição	Proposto	F. Aleatórias	BN-Clássica	n
1	0.4795	0.4797	0.4856	2327
2	0.4799	0.4599	0.4794	2365
3	0.4887	0.4622	0.5057	2357
4	0.4807	0.4840	0.4930	2369
5	0.4942	0.4989	0.5000	2422
6	0.4841	0.4885	0.4962	2394
7	0.4817	0.4968	0.4960	2385
8	0.4676	0.4747	0.4831	2459
Total	0.4820	0.4806	0.4923	19078

É importante salientar que, apesar de termos comparado o modelo de florestas aleatórias com redes Bayesianas, os modelos apresentam possibilidades de aplicação distintas. Enquanto o modelo de florestas aleatórias tem como objetivo tão somente predizer o resultado do processo, os modelos de redes Bayesianas contém a distribuição conjunta de todas as variáveis e, portanto, poderia ser utilizado para predizer qualquer uma das variáveis, considerando diferentes evidências. Nosso modelo também tem algumas vantagens em relação ao modelo de redes Bayesianas clássico, pois lida naturalmente com omissão nas variáveis e possui maior facilidade na inclusão de informações a priori, o que poderia resultar num desempenho melhor do modelo. Os resultados, inclusive, demonstram como pode ser importante utilizar informações à priori nesse tipo de aplicação.

4.4 Decisão

Para tomada de decisão, exploramos a dúvida que o autor de um processo tem ao entrar com a ação. Dadas as informações que o autor tem disponível sobre seu conflito, valeria mais à pena entrar no JEC ou na Justiça Comum e, no caso do JEC, com ou sem um advogado?

Intutitivamente, uma forma de modelar a satisfação é assumir que ficamos satisfeitos quando temos nossa demanda atendida, num intervalo de tempo adequado. Isso significa que maximizaríamos nossa satisfação se o valor de indenização fosse igual ao valor pedido (observe que o valor da indenização geralmente não pode superar o valor do pedido, pois isso configuraria uma decisão *ultra petita*[6]). Ao mesmo tempo, não gostaríamos que o processo demorasse muito. Uma forma de escrever isso em termos de $a =$`valor_acao`, $i =$`resultado_vl` e $t =$`tempo` seria considerar

$$L(t, a, i) = (a - i) * h(t),$$

com h uma função monótona crescente. Uma possível crítica a essa perda seria que esta significaria que não importaria o valor absoluto da indenização, mas somente a relação desse valor com o que foi pedido. Indivíduos que tiverem o interesse em se "aproveitar" da possibilidade de pedir mais, como é o caso da Justiça Comum, não seriam contemplados por essa função. Nesse caso, recomendaríamos a utilização de uma função de perda que não considerasse o valor da ação como, ou então funções que dessem um peso adicional a indenizações de alto valor. Note também que a função L acaba direcionando em certa medida decisões em favor dos JECs pois, intuitivamente, nos JECs o tempo é menor mas também existe uma limitação no valor da indenização, limitação esta que é parcialmente desconsiderada pela função de perda.

Outra característica dessa função de perda é que, no caso de uma indenização completamente satisfatória ($a = i$), teríamos uma perda nula, independentemente do tempo esperado. Uma alternativa neste caso seria utilizar a soma no lugar da multiplicação, por exemplo, com a função $L^*(t, a, i) = (a - i) + h^{'}(t)$.

Lembramos que também realizamos uma transformação nas variáveis para o ajuste dos dados. Sejam $x_a = log(a + 1)$, $x_i = sign(i) \log(|i| + 1)$ e $x_t = \sqrt{t}$. Adaptamos essas variáveis e definimos nossa função de perda final da seguinte forma:

$$L^{**}(t, a, i) = e^{x_a - x_i - \log(2)} + \frac{x_t}{\sqrt{5}}$$

[6] Art. 460 do CPC: É defeso ao juiz proferir sentença, a favor do autor, de natureza diversa da pedida, bem como condenar o réu em quantidade superior ou em objeto diverso do que lhe foi demandado.

Assim, temos que, se $i \geq 0$,

$$L^{**}(t, a, i) = e^{\log(a+1)-\log(i+1)-\log(2)} + \frac{\sqrt{t}}{\sqrt{5}} = \frac{a+1}{2(i+1)} + \sqrt{\frac{t}{5}}.$$

No entanto, se $i < 0$, temos

$$L^{**}(t, a, i) = e^{\log(a+1)-(-\log(-i+1))-\log(2)} + \frac{\sqrt{t}}{\sqrt{5}} = \frac{(a+1)(-i+1)}{2} + \sqrt{\frac{t}{5}}.$$

As constantes que dividem as parcelas da soma foram definidas empiricamente. Convidamos o pesquisador ou pesquisadora interessados a testarem outras funções. O pacote `bnr` possui uma forma de obter as decisões para o modelo, considerando diferentes funções de perda.

Para resolver esse problema, consideramos como informações disponíveis ao indivíduo no momento de sua decisão as variáveis `empresa`, `consumo`, `serasa` e `tipo_dano`. Desconsideramos a informação da gratuidade pois acreditamos que esta esteja mais ligada ao deferimento de gratuidade judiciária pelo juiz do que do pedido de gratuidade em si[7]. Assim, temos que, dados a empresa e, o consumo c, o tipo de dano d e colocação do nome no Serasa s, queremos encontrar o tipo de vara v^* que minimiza a perda esperada em relação à perda L^{**}. Em símbolos, temos

$$v^* = \text{argmin}_v \mathbb{E}\Big[L^{**}\Big|\mathbf{X}; e, c, s, d, v\Big].$$

O valor esperado pode ser escrito como

$$\mathbb{E}\Big[L^{**}\Big|\mathbf{X}; e, c, s, d, v\Big] = \int L^{**}(t, a, i)p(t, a, i|\mathbf{X}; e, c, s, d, v).$$

A função de densidade p pode ser trabalhada com base na estrutura definida pela rede Bayesiana. Nas operações que seguem, vamos omitir os dados $\mathbf{X}$ por conveniência. Primeiramente, utilizamos a probabilidade total em relação aos possíveis valores do resultado r, e utilizamos as regras da probabilidade condicional nas variáveis, já desconsiderando algumas variáveis por conta das independências condicionais:

$$p(t, a, i|e, c, s, d, v) = \sum_r p(t|r, d, v) * p(i|a, r)p(a|e, d, v)p(r|e, c, s, d, v).$$

Note que, neste ponto, a única expressão que ainda não tem distribuição definida (dado $\boldsymbol{\theta}$) é $p(r|e, c, s, d, v)$. Podemos resolver aplicando novamente a probabilidade total em relação à variável gra-

[7] Não podemos afirmar com certeza pois a informação foi extraída dos textos e a palavra aparece tanto quando pedimos gratuidade quanto quando o juiz a defere.

tuidade g, obtendo

$$p(t,a,i|e,c,s,d,v) = \sum_r p(t|r,d,v) * p(i|a,r)p(a|e,d,v) \sum_g p(r|g,e,c,s,d,v)p(g).$$

Para obtenção dos valores em cada p, a posteriori, aplicamos as distribuições preditivas como mencionado anteriormente. A integral é obtida através de métodos numéricos. No nosso caso, utilizamos o pacote `cubature` para a tarefa. Para cada uma das 36 combinações de `empresa`, `consumo`, `serasa` e `tipo_dano`, obtivemos a perda esperada para cada `tipo_vara` e identificamos a que apresentava a menor perda.

Os resultados podem ser observados nas Tabelas 4.2, 4.3 e 4.4, uma para cada empresa. Observe que, na maioria dos casos, temos como decisão ótima o JEC, com ou sem advogado. A decisão fica na Justiça Comum somente em alguns casos específicos, especialmente quando envolvem bancos e danos materiais.

Tabela 4.2: *decisão ótima, perda esperada e erro estimado no cálculo do valor esperado, para cada configuração de consumo, serasa e tipo de dano, quando a empresa é um banco.*

empresa	consumo	serasa	tipo_dano	decisao	L_esti	L_erro
BANCO	não	não	dano material	Comum com advogado	15.095	0.053
BANCO	não	não	dano moral	JEC com advogado	25.794	0.089
BANCO	não	não	dano moral e material	JEC com advogado	25.839	0.054
BANCO	não	sim	dano material	JEC sem advogado	11.403	0.058
BANCO	não	sim	dano moral	JEC sem advogado	16.771	0.047
BANCO	não	sim	dano moral e material	JEC com advogado	17.332	0.058
BANCO	sim	não	dano material	Comum com advogado	13.176	0.019
BANCO	sim	não	dano moral	JEC sem advogado	25.466	0.044
BANCO	sim	não	dano moral e material	JEC com advogado	17.699	0.029
BANCO	sim	sim	dano material	JEC com advogado	14.391	0.032
BANCO	sim	sim	dano moral	JEC com advogado	12.410	0.039
BANCO	sim	sim	dano moral e material	Comum com advogado	13.338	0.029

Tabela 4.3: *decisão ótima, perda esperada e erro estimado no cálculo do valor esperado, para cada configuração de consumo, serasa e tipo de dano, quando a empresa é de telefonia móvel.*

empresa	consumo	serasa	tipo_dano	decisao	L_esti	L_erro
TEL.	não	não	dano material	JEC com advogado	5.397	0.037
TEL.	não	não	dano moral	JEC com advogado	10.538	0.048
TEL.	não	não	dano moral e material	JEC com advogado	8.348	0.026
TEL.	não	sim	dano material	JEC sem advogado	5.886	0.021
TEL.	não	sim	dano moral	JEC sem advogado	8.117	0.016
TEL.	não	sim	dano moral e material	JEC sem advogado	6.492	0.011
TEL.	sim	não	dano material	JEC sem advogado	6.665	0.014

Tabela 4.3: *decisão ótima, perda esperada e erro estimado no cálculo do valor esperado, para cada configuração de consumo, serasa e tipo de dano, quando a empresa é de telefonia móvel.*

empresa	consumo	serasa	tipo_dano	decisao	L_esti	L_erro
TEL.	sim	não	dano moral	JEC sem advogado	9.514	0.019
TEL.	sim	não	dano moral e material	JEC com advogado	6.980	0.023
TEL.	sim	sim	dano material	JEC sem advogado	6.741	0.024
TEL.	sim	sim	dano moral	JEC com advogado	6.601	0.009
TEL.	sim	sim	dano moral e material	JEC com advogado	5.812	0.008

Tabela 4.4: *decisão ótima, perda esperada e erro estimado no cálculo do valor esperado, para cada configuração de consumo, serasa e tipo de dano, quando a empresa é NET ou Eletropaulo.*

empresa	consumo	serasa	tipo_dano	decisao	L_esti	L_erro
OUTROS	não	não	dano material	JEC com advogado	8.122	0.058
OUTROS	não	não	dano moral	JEC com advogado	14.687	0.065
OUTROS	não	não	dano moral e material	JEC com advogado	23.470	0.089
OUTROS	não	sim	dano material	JEC sem advogado	9.178	0.042
OUTROS	não	sim	dano moral	JEC sem advogado	6.862	0.028
OUTROS	não	sim	dano moral e material	JEC sem advogado	16.36	0.091
OUTROS	sim	não	dano material	JEC sem advogado	10.730	0.025
OUTROS	sim	não	dano moral	JEC sem advogado	11.397	0.035
OUTROS	sim	não	dano moral e material	JEC com advogado	17.692	0.062
OUTROS	sim	sim	dano material	JEC com advogado	11.782	0.042
OUTROS	sim	sim	dano moral	JEC sem advogado	7.399	0.020
OUTROS	sim	sim	dano moral e material	JEC com advogado	14.292	0.039

No caso em que temos "outras empresas", relação de consumo, colocação de nome no Serasa e dano moral e material, a decisão ótima seria entrar no JEC com advogado.

Capítulo 5

Considerações finais

"Done is better than perfect."

Paredes nos prédios da sede do Facebook.

Neste trabalho exploramos diversos conceitos novos, que usualmente não são objeto de estudo da graduação em estatística. Os modelos gráficos mostraram-se bastante flexíveis e facilitaram o alinhamento de conhecimento de especialistas com inferência estatística, sem com isso perder o poder preditivo. Além disso, a abordagem Bayesiana se mostrou sem dificuldades técnicas, pois atualmente é fácil superar seus problemas computacionais para problemas de pequeno e médio porte.

Em relação à pesquisa, tivemos resultados úteis para decidir entre ir à justiça com processos nos JECs ou na Justiça Comum. No caso deste autor, por exemplo, o ideal seria entrar no JEC com advogado. Advogados poderiam utilizar as tabelas como parâmetros para entrada de processos de seus clientes.

Em relação à parte computacional, é importante mencionar que negligenciar a tarefa de extração e manipulação de dados como parte da estatística é ruim para nossa profissão. Somente neste projeto, estimamos que a tarefa de extração e consolidação da base tomou 90% do tempo total de execução. Acreditamos que a academia deveria focar mais em produções que auxiliem pesquisadores a trabalharem com os conceitos de extração, manipulação e visualização dos dados com mais cuidado, profundidade e eficiência.

Sobre a jurimetria, ainda temos um bebê, mas ele já está começando a dar seus próprios passos. Recentemente, temos visto muitos alunos da estatística e do direito (ou de ambos) interessando-se no tema, o que recebemos com muita felicidade. Esperamos que este trabalho possa funcionar como um guia para pesquisas *ad-hoc* em jurimetria, e ficaríamos muito satisfeitos em discutir sobre os conceitos de jurimetria, as dificuldades técnicas e melhores abordagens nessa área.

Sobre reproditibilidade, esperamos que o leitor tenha notado nosso interesse em tornar nosso estudo reprodutível. O trabalho é complexo, mas isso faz parte de algo que acreditamos ser positivo para a

produção científica. Além disso, mais uma vez, encorajamos os pesquisadores de estatística a terem uma preocupação especial nesse ponto, produzindo não só a base matemática para os modelos, como também pacotes *open-source* que possibilitem sua utilização.

Acreditamos que um passo positivo foi dado. Com nossos esforços para distribuir os códigos e desenvolver uma aplicação do início ao fim, desde a obtenção de dados até a tomada de decisões, foi possível preparar o terreno para pesquisadores que desejarem realizar estudos mais avançados.

5.1 Pesquisas futuras

Praticamente toda pesquisa gera mais perguntas do que respostas. Nosso intuito neste trabalho não foi fechar os tópicos discutidos e, por isso, convidamos pesquisadores interessados nos temas de diagramas de influências e jurimetria a criticarem e sugerirem novas abordagens.

Abaixo, levantamos alguns tópicos que não foram abordados na pesquisa, mas que acreditamos serem interessantes e promissores:

- Modelos "tópicos" para análise de textos. Na discussão sobre os dados, mencionamos que utilizamos modelos estatísticos para auxiliar na classificação de algumas variáveis baseadas no texto. No final, acabamos subutilizando o poder dos *topic models*, e seria algo que poderia melhorar a análise dos textos.
- Imagine que agora, a partir de um modelo construído, podemos criar rotinas computacionais que extraiam automaticamente processos da web e produzam predições para seus resultados. Seria interessante tentar construir algo nesse sentido. Isso poderia, posteriormente, se tornar um produto útil no direito.
- Outra linha de interesse seria desconstruir os litígios de forma mais teórica e realizar a elicitação das prioris com cuidado. Isso poderia produzir resultados muito melhores com base em informações de especialistas, e colaborariam na construção de fundamentos da jurimetria.
- Seria interessante também reproduzir esse estudo aplicado a outros temas jurídicos de interesse, e testar novas maneiras de obtenção de dados dos tribunais. As recentes iniciativas sobre processos digitais podem ajudar bastante o pesquisador nesse sentido.

Apêndice A

Redes Bayesianas

As redes Bayesianas são o resultado da combinação de conceitos probabilísticos e conceitos da teoria dos grafos. Segundo Pearl (2009), tal união tem como consequências três benefícios: i) prover formas convenientes para expressar suposições do modelo; ii) facilitar a representação de funções de probabilidade conjuntas; e iii) facilitar o cálculo eficiente de inferências a partir de observações.

A.1 Probabilidade condicional

Da teoria de probabilidades precisamos apenas de alguns resultados básicos sobre probabilidade condicional. Primeiramente, pela definição de probabilidade condicional, sabemos que

$$p(x_1, x_2) = p(x_1)p(x_2|x_1).$$

Aplicando essa regra iterativamente para n variáveis, temos

$$p(x_1, \ldots, x_p) = \prod_j p(x_j \mid x_1, \ldots, x_{j-1}).$$

Agora, imagine que, no seu problema, a variável aleatória X_j não dependa probabilisticamente de todas as variáveis $X_1, \ldots, X_{j-1}$, e sim apenas de um subconjunto Π_j dessas variáveis. Fazendo isso, a equação pode ser escrita como

$$p(x_1, \ldots, x_p) = \prod_j p(x_j \mid \pi_j).$$

As suposições de independência condicional são o grande diferencial das redes Bayesianas. Com elas, é possível reduzir consideravelmente o número de parâmetros e ainda ter acesso a toda a distribuição de probabilidades.

Ao conjunto de variáveis aleatórias Π_j damos o nome de **pais markovianos** de X_j, ou simplesmente

pais de X_j. A escolha desse termo ficará mais clara quando tratarmos de grafos. O conjunto Π_j pode ser pensado também como o conjunto de informações suficientes para determinar as probabilidades de X_j.

De início, pode parecer pouco prático definir um conjunto de pais Π_j para cada variável X_j. Como veremos em seguida, no entanto, é possível criar uma forma gráfica que resume toda essa informação de forma eficiente.

A.2 Grafos

Um **grafo direcionado** é um par ordenado $G = (V, E)$, em que cada elemento de E corresponde a um par não-ordenado de elementos de V. Os elementos do conjunto V são chamados de **vértices** e os elementos de E são chamados de **arestas**. Adicionando a restrição de ordenamento dos elementos de V em cada aresta, definimos um **grafo direcionado**. Nesse caso, chamamos os elementos de E de **setas**.

Uma **trilha** em um grafo direcionado é uma sequência finita $P = (v_0, e_1, v_1, e_2, \ldots, e_k, v_k)$, cujos termos são alternadamente vértices v_i e setas e_j e tal que, para todo i, $1 \leq i \leq k$, os extremos de e_i sao v_{i-1} e v_i.

Uma trilha de comprimento não nulo (ou seja, com pelo menos uma seta) com origem e término coincidentes e tal que todos os vértices são distintos é denominada **circuito**. Um grafo é considerado **acíclico** se não contém circuitos.

Um grafo direcionado é **simples** se não tem laços (setas com extremos iguais) e nem setas múltiplas (setas com os mesmos extremos). No nosso caso, temos interesse em um tipo muito específico de grafo, que é o grafo direcionado simples e acíclico, ou **DAG** (*directed acyclic graph*).

Exemplo 1. *O chão escorregadio* (Pearl, 2009). Na Figura A.1 temos uma situação envolvendo a estação do ano, o fato de estar chovendo ou não, regador ligado ou desligado, o chão estar molhado e o chão estar escorregadio. É possível gerar o gráfico utilizando a função `grViz` do pacote `DiagrammeR` do R.

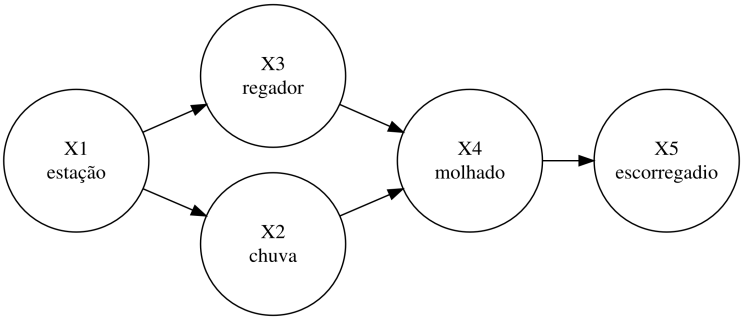

Figura A.1: *Exemplo de DAG.*

A.3 Unindo os conceitos

A união dos dois conceitos supracitados se dá através teorema mostrando que, se temos um conjunto $X_1, \ldots, X_p$ e seus pais $\Pi_1, \ldots, \Pi_j$, é possível construir um único DAG em que cada seta $X_k \to X_j$ ocorre se e só se $X_k \in \Pi_j$. Ou seja, os pais de X_j, no grafo, são os nós que apontam para X_j[1].[2]

Definição (Rede Bayesiana). Uma rede Bayesiana é uma tripla

$$(G, X, P)$$

em que

a) $G = (V, E)$ é um grafo acíclico direcionado (DAG) com vértices V e arestas E.

b) X é um conjunto de variáveis aleatórias em que cada variável X_i é representada por um vértice $v_i \in V$, e cada dependência condicional é representada por uma seta $e \in E, e = (v_k, v_j)$, com v_k representando $X_k \in \Pi_j$ e v_j representando X_j.

c) P é um conjunto de funções de probabilidade condicionais compatível com G.

Exemplo 2. Continuando no exemplo do chão escorregadio, o DAG da Figura 1 representa a seguinte distribuição de probabilidades

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_4).$$

Vamos assumir, neste momento, que todas as variáveis $X_1, \ldots, X_p$ são categóricas. Temos então que cada X_j tem distribuição multinomial. Assim, se X_j tem s_j categorias, temos um conjunto θ_j com $s_j - 1$ parâmetros (o s_j-ésimo é dado por um menos a soma dos anteriores), para cada possível configuração de π_j.

Além disso, temos que $\sum_{k=1}^{s_j} x_k = n_j$. Em problemas mais comuns, temos $n_j = 1$, ou seja, cada X_j é uma bernoulli multivariada. Nas expressões que seguem, vamos assumir que $n_j = 1$ para simplificar a notação, e convidamos o leitor a construir os mesmos resultados no caso multinomial.

Podemos escrever a distribuição de probabilidades condicional de X_j dado π_j e θ_{π_j} da seguinte forma:

[1] É possível mostrar também que, se temos um DAG representando todas as variáveis $X_1, \ldots, X_p$ e se $p(x_1, \ldots, x_p) > 0$ para toda combinação de $x_1, \ldots, x_p$, então o conjunto $\Pi_1, \ldots, \Pi_n$ é único.

[2] É importante mencionar também que, a partir de um conjunto $X_1, \ldots, X_p$ e seus pais $\Pi_1, \ldots, \Pi_j$, podem existir diversos DAGs que representam a mesma distribuição de probabilidades conjunta $p(x_1, \ldots, x_p)$. Esses DAGs podem ser construídos através de técnicas de inversão de setas, que podem ser interessantes computacionalmente. Existe também um critério, chamado critério de d-separação, que pode ser utilizado para gerar todos os possíveis DAGs que geram a mesma P.

$$p(x_k \mid \pi_j, \theta_{\pi_j}) = \prod_{k}^{s_j} (\theta_{x_k|\pi_j})^{x_{ik}}, \text{com} \sum_{k}^{s_j} x_{ik} = 1.$$

Agora, digamos que temos um estudo com uma base de dados com n observações $X_i = (X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$. A função de log-verossimilhança é dada por

$$l(X, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{\pi_j} \log p(x_{ij}|\pi_j, \theta_j),$$

em que

$$\log p(x_{ij}|\pi_j, \theta_j) = \sum_{k}^{s_j} \theta_{x_k|\pi_j} x_{ik}, \text{ com } \sum_{k}^{s_j} x_{ik} = 1.$$

Em um cenário frequentista, é possível obter os estimadores de máxima verossimilhança de forma trivial, através do cálculo das frequências

$$\hat{\theta}_{x_{jk}|\pi_j} = \frac{N(x_{jk}, \pi_j)}{N(\pi_j)},$$

em que $N(x_k, \pi_j)$ é o número de observações com $X_j = x_k$ e $\Pi_j = \pi_j$ e $N(\pi_j)$ é o número de observações com $\Pi_j = \pi_j$.

Em um cenário Bayesiano, o problema pode complicar consideravelmente pois temos maior liberdade para definir prioris para os parâmetros. Por exemplo, poderíamos considerar que os parâmetros para duas configurações possíveis os pais de X_j sejam dependentes.

A abordagem mais "ad-hoc" para atribuir as posterioris é supor que θ_{π_j}, parâmetros de $X_j|\pi_j$ têm distribuição Dirichlet, e que θ_{π_j} é independente de $\theta_{\pi_j'}$, se $\pi_j' \neq \pi_j$. Essa forma poderia ser visualizada aumentando-se o DAG original, adicionando um nó θ_{π_j} apontando para X_j para cada configuração de π_j. Uma das vantagens dessa forma de inserir as prioris é que as posterioris são conjungadas e, assim, se a priori for dada por

$$\theta_{\pi_j} \sim \text{Dir}(\alpha_{1;\pi_j}, \ldots, \alpha_{s_j;\pi_j}),$$

a posteriori fica

$$\theta_{\pi_j}|X \sim \text{Dir}(\alpha_{1;\pi_j} + c_{j1}, \ldots, \alpha_{s_j;\pi_j} + c_{js_j}),$$

em que $c_l = \sum_i x_{ijl}$

A.4 Trabalhando com omissão nas variáveis

Quando um estudo envolve informações faltantes, a primeira preocupação que devemos ter é se existe algum mecanismo para a produção dos dados faltantes, e se esse mecanismo estaria relacionado com os dados observados ou não observados. Ignorar dados faltantes pode tornar uma pesquisa inviável, por conter poucas informações incompletas, e também pode gerar viés nos resultados.

Os fundamentos sobre omissão nas variáveis podem ser encontrados em Little e Rubin (2014). Podemos classificar os dados faltantes em três tipos principais:

- Missing completamente aleatório (MCAR). Não depende de nenhuma outra variável.
- Missing aleatório (MAR). A probabilidade de missing depende somente das informações observadas.
- Missing não aleatório (MNAR). Outros casos.

Nesse texto, vamos lidar apenas com MCAR e MAR. Nesses casos, o mecanismo gerador dos dados faltantes pode ser considerado *ignorável*, o que facilita nossas análises. Na prática, isso significa que o mecanismo gerador dos dados faltantes depende somente de dados observáveis, e podemos tratar os dados faltantes como novos parâmetros a serem estimados no modelo.

Uma vantagem da abordagem Bayesiana é que trabalhar com dados faltantes torna-se algo natural. Seja $x_i = z_i, y_i$ uma observação, de forma que $z_i \subset x_i$ é o conjunto de dados observados e $y_i = x_i \; z_i$ é o conjunto de dados omissos. Nesse caso, poderíamos calcular a posteriori da seguinte maneira:

$$p(\theta_{\pi_j}|z_i) = \sum_{y_i} p(\theta_{\pi_j}|y_i, z_i)$$

É possível mostrar que, nesse caso, temos que a distribuição da posteriori é uma *mistura de Dirichlets* (Spiegelhalter e Lauritzen (1990)). Se encontrarmos mais observações com dados faltantes, teremos misturas das misturas de dirichlets obtidas previamente. Para bases de dados com muitos dados faltantes, isso pode ficar computacionalmente intratável e, por isso, precisamos recorrer a métodos computacionais.

O método computacional utilizado neste trabalho é o amostrador de Gibbs (Geman e Geman (1984)), um método Monte Carlo que permite gerar amostras da distribuição à posteriori após um número suficientemente grande de iterações.

O algoritmo para gerar o amostrador de Gibbs neste caso seria

1. Começamos com um conjunto de dados incompleto, uma rede Bayesiana e um conjunto de prioris Dirichlet independentes.
2. Completamos a base de dados aleatoriamente.

3. Escolhemos uma variável que não foi observada e imputamos esse valor com a preditiva. Repetimos esse passo até imputar todas as variáveis.
4. Atualizamos os parâmetros a partir do banco de dados imputado. Volta ao segundo passo.

A.4.1 Trabalhando com variáveis contínuas

No paradigma Bayesiano, incluir variáveis contínuas também não é necessariamente complexo. Para isso, no entanto, seria preciso que nenhuma variável discreta tenha como pai uma variável contínua, que as prioris para as variáveis contínuas sejam independentes das prioris para as variáveis discretas e que a distribuição da variável contínua seja tal que a posteriori é conjugada.

No caso em que temos uma variável discreta como pai da variável contínua, seria necessário definir uma função de ligação para associar o valor da variável contínua com os parâmetros da multinomial[3]. Por exemplo, digamos que temos $X|\theta \sim Bernoulli(\theta)$, condicionalmente dependente de $Y|\mu, \sigma \sim N(\mu, \sigma)$. Podemos definir uma função de ligação logística e considerar os parâmetros α e β, de modo que

$$P(X = 1|y, \alpha, \beta) = \frac{1}{1 + e^{-\alpha - \beta y}}.$$

Nesses casos, a parte computacional pode ser dificultada. Ainda existe a possibilidade de definição de prioris conjugadas (Chen e Ibrahim (2003)), mas estas nem sempre são fáceis de interpretar e, usualmente, acabamos partindo para métodos MCMC mais gerais, como o algoritmo de Metropolis (Hastings (1970)).

Em nossa aplicação não temos nenhum exemplo de variável discreta com pais contínuos. Dessa forma, o amostrador de Gibbs será suficiente para atingir nossos objetivos.

Cálculos de probabilidades com redes Bayesianas Uma das principais aplicações quando construímos redes Bayesianas é a possibilidade de calcular probabilidades de acordo com uma *consulta*. Uma consulta consiste na avaliação de uma probabilidade do tipo

$$P(\tilde{X} = \tilde{x}|\mathbf{E} = \mathbf{e}),$$

em que $\tilde{X}$ é um conjunto de variáveis de interesse e $\mathbf{E}$ é um conjunto de evidências. A partir uma rede Bayesiana completa, é possível avaliar qualquer tipo de consulta, já que temos uma expressão para a distribuição conjunta de todas as variáveis. Na prática, no entanto, encontramos algumas dificuldades.

Em geral, a avaliação de uma consulta em uma rede Bayesiana é um problema computacionalmente complexo. Isso acontece porque, dependendo da consulta, precisamos aplicar o teorema da probabilidade

[3] Ver Murphy (1999) para detalhes.

total múltiplas vezes para obter a expressão em função das probabilidades condicionais conhecidas. Para redes com variáveis contínuas, isso pode incluir integrações que podem ter custo computacional elevado. Nesses casos, é comum a aplicação de métodos de integração numérica e métodos de Monte Carlo para as avaliações.

Uma alternativa conveniente para avaliação das probabilidades de eventos é construir a rede de forma "causal". Dessa forma, as avaliações que precisamos fazer na prática, do tipo "efeito dado causa", ou "depois dado antes", são também aquelas que possuem avaliação mais simples, pela própria construção da rede.

Na abordagem Bayesiana, a obtenção das probabilidades da rede à posteriori também não é necessariamente direta. Para isso, utilizamos a *distribuição preditiva* da posteriori. Suponha que nosso interesse seja avaliar a probabilidade à posteriori de $\tilde{X}$ dado $\mathbf{E} = \mathbf{e}$, ou seja,

$$p(\tilde{X} = \tilde{x}|\mathbf{X}; \mathbf{E} = \mathbf{e})$$

Note que que, nesse contexto, $\tilde{X}$ e $\mathbf{X}$ não são independentes, mas sim condicionalmente independentes dado o vetor de parâmetros $\boldsymbol{\theta}$. Utilizando o teorema da probabilidade total, temos

$$p(\tilde{X} = \tilde{x}|\mathbf{X}; \mathbf{E} = \mathbf{e}) = \int_{\boldsymbol{\theta}} p(\tilde{X} = \tilde{x}|\mathbf{X}; \boldsymbol{\theta}, \mathbf{E} = \mathbf{e}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} p(\tilde{X} = \tilde{x}|\boldsymbol{\theta}; \mathbf{E} = \mathbf{e}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta}.$$

Observe que $p(\tilde{X} = \tilde{x}|\boldsymbol{\theta}; \mathbf{E} = \mathbf{e})$ pode ser avaliado pelas distribuições definidas na rede Bayesiana. Assim, temos que a probabilidade desejada é o valor esperado à posteriori

$$p(\tilde{X} = \tilde{x}|\mathbf{X}; \mathbf{E} = \mathbf{e}) = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}}\Big[p(\tilde{X} = \tilde{x}|\boldsymbol{\theta}; \mathbf{E} = \mathbf{e})\Big].$$

Essas quantidades podem ser obtidas diretamente caso as prioris sejam conjugadas, através das distribuições preditivas. Por exemplo, no caso da priori Dirichlet, temos uma distribuição preditiva Dirichlet-multinomial e, no caso da priori normal-gama-inversa, temos uma distribuição preditiva t-student não central (Griffin *et al.* (2010)).

Se temos uma amostra da posteriori $\theta|\mathbf{X}$, é possível estimar a probabilidade calculando-se a média das probabilidades avaliadas para cada elemento da amostra. Esse método pode ser computacionalmente custoso se a referida probabilidade for difícil de avaliar.

Apêndice B

Pacotes utilizados

Neste apêndice, vamos mostrar, de forma simplificada, como os pacotes `tjsp` e `bnr` podem ser utilizados para extração de dados do TJSP e para ajuste de modelos de redes bayesianas. As descrições não são completas nem definitivas, pois, até a publicação dos pacotes no CRAN (caso isso venha a acontecer), provavelmente algumas funções terão sofrido mudanças.

Vale notar que não foram somente estes pacotes que desenvolvemos no R. A saber, os outros pacotes são `jmasters`, que contém o texto da dissertação de mestrado; `mcmc`, que contém diversos testes relacionados a algoritmos mcmc que serão, posteriormente, adicionados ao pacote `bnr`; `amostragem`, que possui algumas funções específicas para download de dados dos tribunais via amostragem; `tjsp.data`, que contém scripts para limpeza dos dados; `prevproc`, que é um pacote, ainda em fase embrionária, cujo objetivo é prever os resultados de um processo, dado seu número; `influenced`, que também está em fase embrionária, mas traria uma API para utilização de diagramas de influências e tomadas de decisão, a ser utilizado em conjunto com o pacote `bnr`; e `thesis`, um pacote que permite a escrita de uma dissertação de mestrado com o modelo do IME-USP, baseado em RMarkdown.

B.1 Pacote `tjsp`

O objetivo do pacote `tjsp` é simplesmente facilitar a obtenção de dados a partir do TJSP. O pacote é feito inteiramente em R e possui funções para extração de dados da Consulta de Julgados de Primeiro Grau (CJPG, consulta de sentenças), Consulta de Julgados de Segundo Grau (CJSG, consulta de acórdãos), Consulta de Processos de Primeiro Grau (CPO-PG), Consulta de Processos de Segundo Grau (CPO-SG), Diário Oficial Eletrônico (DOE) e peças processuais. O pacote também apresenta facilidades para geração aleatória de números de processos com base no padrão CNJ.

B.1.1 Instalação

O pacote `tjsp` não está disponível no CRAN. Para instalá-lo, é necessário antes obter o pacote `devtools`.

```
if(!require(devtools)) install.packages('devtools')
devtools::install_github('jtrecenti/tjsp')
```

B.1.2 Utilização

Vamos demonstrar a utilização das funções `cjpg` e `cpopg`. A primeira é útil para baixar sentenças a partir de informações para consulta, e a segunda para baixar processos específicos a partir de seus números.

Os parâmetros da função `cjpg` são

- `livre`: palavras-chave a serem pesquisadas.
- `classes`: códigos das classes processuais.
- `assuntos`: códigos dos assuntos.
- `data_inicial`: data inicial da publicação da sentença.
- `data_final`: data final da publicação da sentença
- `varas`: códigos das varas.
- `min_pag`: primeira página do resultado a ser pesquisada (cada página contém 10 sentenças),
- `max_pax`: última página do resultado. `max_pax=Inf` baixará todas as páginas.
- `salvar`: informa se os arquivos html devem ser salvos em disco.
- `path`: informa a pasta em que os arquivos html serão salvos.

Exemplo: Vamos baixar a primeira página de resultados usando como termo livre "lucros cessantes".

```
lc <- tjsp::cjpg(livre = 'lucros cessantes', max_pag = 1L)
```

```
## pag: 1...downloading...download realizado! parsing...OK!
```

```
lc
```

```
## Source: local data frame [10 x 11]
##
##                    classe                              assunto
##                     (chr)                                (chr)
## 1  Procedimento Ordinário          Indenização por Dano Moral
```

```
## 2  Procedimento Ordinário       Indenização por Dano Material
## 3    Procedimento Sumário                Acidente de Trânsito
## 4  Procedimento Ordinário       Indenização por Dano Material
## 5  Procedimento Ordinário Interpretação / Revisão de Contrato
## 6  Procedimento Ordinário       Indenização por Dano Material
## 7  Procedimento Ordinário       Indenização por Dano Material
## 8  Procedimento Ordinário               Espécies de Contratos
## 9  Procedimento Ordinário                      Inadimplemento
## 10   Procedimento Sumário                           Pagamento
## Variables not shown: magistrado (chr), comarca (chr), foro (chr), vara
##   (chr), data_de_disponibilizacao (chr), n_processo (chr), cod_sentenca
##   (chr), txt (chr), pag (int)
```

O resultado é um `data.frame` com 10 processos, contendo a informação da classe, assunto, magistrado, comarca, foro, vara, data de disponibilização, número do processo, código da sentença, texto da sentença e página.

Digamos que, a partir dos processos obtidos em `lc`, desejamos obter mais informações, através do CPO-PG. podemos fazer isso com as funções `cpopg` e `parse_cpopg`:

```
temp_dir <- 'temp' # pasta temporaria para guardar os arquivos
dir.create(temp_dir)
```

```
## Warning in dir.create(temp_dir): 'temp' already exists
```

```
tjsp::cpopg(lc$n_processo, temp_dir)

arqs <- dir(temp_dir, full.names = TRUE)
lc_cpopg <- tjsp::parse_cpopg(arqs)
lc_cpopg
```

```
## Source: local data frame [10 x 4]
##
##                              arq           infos         partes
##                            (chr)          (list)         (list)
## 1  temp/00014963720098260200.html <tbl_df [14,2]> <tbl_df [8,3]>
## 2  temp/00108601720148260084.html <tbl_df [14,2]> <tbl_df [4,3]>
## 3  temp/00469484520098260564.html <tbl_df [14,2]> <tbl_df [2,3]>
## 4  temp/10026087220158260019.html <tbl_df [13,2]> <tbl_df [4,3]>
## 5  temp/10059435520138260606.html <tbl_df [13,2]> <tbl_df [4,3]>
## 6  temp/10078935020148260320.html <tbl_df [12,2]> <tbl_df [7,3]>
## 7  temp/10163939220148260001.html <tbl_df [13,2]> <tbl_df [6,3]>
## 8  temp/10391308320148260100.html <tbl_df [12,2]> <tbl_df [2,3]>
```

```
## 9  temp/30002477720138260614.html <tbl_df [13,2]> <tbl_df [7,3]>
## 10 temp/30003652220138260495.html <tbl_df [13,2]> <tbl_df [7,3]>
## Variables not shown: movs (chr)
```

A função `cpopg` simplesmente baixa os arquivos HTML e salva na pasta selecionada. A função `parse_cpopg` lê os arquivos HTML e retorna um `data.frame` com quatro colunas: nome do arquivo, informações, partes e movimentações, sendo que cada um dos três últimos também é um `data.frame`. As informações de uma fonte de dados podem ser expandidas com o seguinte código:

```
# pega as informações
library(dplyr)
library(tidyr)

lc_cpopg_infos <- lc_cpopg %>%
  select(arq, infos) %>%
  unnest(infos)

lc_cpopg_infos
```

```
## Source: local data frame [131 x 3]
##
##                               arq            key
##                             (chr)          (chr)
## 1  temp/00014963720098260200.html       processo
## 2  temp/00014963720098260200.html         classe
## 3  temp/00014963720098260200.html           area
## 4  temp/00014963720098260200.html        assunto
## 5  temp/00014963720098260200.html   local_fisico
## 6  temp/00014963720098260200.html   distribuicao
## 7  temp/00014963720098260200.html          lugar
## 8  temp/00014963720098260200.html           juiz
## 9  temp/00014963720098260200.html outros_numeros
## 10 temp/00014963720098260200.html  valor_da_acao
## .. ...                                      ...
## Variables not shown: value (chr)
```

B.2 Pacote `bnr`

O pacote `bnr` tem o singelo objetivo de ajustar os parâmetros de redes bayesianas baseado num amostrador de gibbs. No pacote, assumimos que as variáveis discretas têm distribuição multinomial com priori Dirichlet, e que as variáveis contínuas têm distribuição normal com priori normal-gama-inversa.

A vantagem do `bnr` em relação a outros pacotes semelhantes como `bnlearn` e `deal` é que podemos trabalhar com omissão nas variáveis de forma direta. A utilização do pacote `bnr` para bases de dados completas irá simplesmente gerar amostras de uma distribuição a posteriori com forma analítica definida e, portanto, pode ser menos útil do que o pacote `deal`, por exemplo. Além disso, somente trabalhamos os problemas pela abordagem bayesiana então, caso o pesquisador tenha interesse em ajustar um modelo usando estatística clássica, recomendamos utilizar o pacote `bnlearn`. O pacote `bnr` também não trabalha, até o momento, com aprendizado da estrutura da rede.

O ajuste do modelo pode ser feito realizando-se k partições aleatórias da base de dados (*k-folds*). Isso pode ser útil caso o usuário tenha interesse em avaliar o poder preditivo do modelo.

B.2.1 Instalação

O pacote `bnr` não está disponível no CRAN. Para instalá-lo, é necessário antes obter o pacote `devtools`.

```
if(!require(devtools)) install.packages('devtools')
devtools::install_github('jtrecenti/bnr')
```

B.2.2 Utilização

Um modelo de redes bayesianas precisa de dois elementos principais: a estrutura da rede e os dados. Para definir a estrutura da rede, fazemos

```
edges <- list(
  'resultado' = 'resultado_vl',
  'tipo_dano' = 'valor_acao',
  'valor_acao' = 'resultado_vl',
  'gratuidade' = 'resultado',
  'consumo' = 'resultado',
  'serasa' = 'resultado',
  'empresa' = 'serasa',
  'tipo_vara' = 'valor_acao',
  'empresa' = 'valor_acao',
  'empresa' = 'resultado',
  'tipo_vara' = 'resultado',
  'tipo_vara' = 'tempo',
  'tipo_dano' = 'tempo',
  'empresa' = 'consumo',
  'resultado' = 'tempo',
  'tipo_dano' = 'resultado'
```

```
)
bn <- bnr::create_bn(edges)
```

A função `create_bn` é apenas um atalho para as funções do pacote `bnlearn`. Veja `help('bn class', package = 'bnlearn')` para detalhes.

Os dados devem estar armazenados em um `data.frame` com os nomes das variáveis iguais aos nomes dos nós da rede bayesiana. Variáveis discretas devem ser da classe `character` e variáveis contínuas devem ser da classe `numeric`.

```
library(dplyr) # facilita a visualização do data.frame
data('d_cjpg', package = 'bnr')
d_cjpg
```

```
## Source: local data frame [19,078 x 10]
##
##      empresa          tipo_vara valor_acao resultado_vl serasa
##        (chr)              (chr)      (dbl)        (dbl)  (chr)
## 1      BANCO   JEC sem advogado 0.17103664   0.17103664    não
## 2     OUTROS   JEC sem advogado 0.06385997           NA    não
## 3  TELEFONIA   JEC sem advogado 0.06385997           NA    sim
## 4      BANCO Comum com advogado 0.40770460           NA    sim
## 5      BANCO   JEC sem advogado 0.11790696   0.11790696    não
## 6  TELEFONIA   JEC sem advogado 0.02152109   0.02152109    sim
## 7      BANCO Comum com advogado 0.08957922  -0.08957922    não
## 8  TELEFONIA   JEC sem advogado 0.03388328   0.03388328    não
## 9  TELEFONIA   JEC sem advogado 0.06081540   0.06081540    não
## 10     BANCO Comum com advogado 0.08957922   0.08957922    não
## ..       ...               ...        ...          ...    ...
## Variables not shown: resultado (chr), gratuidade (chr), tempo (dbl),
##   tipo_dano (chr), consumo (chr)
```

A função `bnr` tem como objetivo estruturar os dados de forma adequada para realização da amostragem. Essa funções faz as computações necessárias (e.g. as tabelas de proporções condicionais, modelos lineares, etc.) para o ajuste do modelo. Essa função também permite que a base seja particionada em k partições, para que seja possível realizar validações. Se o usuário não desejar realizar partições, basta utilizar `k=1` (o valor padrão).

```
dados_bn <- bnr::bnr(d_cjpg, edges, kfolds = 8L)
```

Para ajuste do modelo, utilizamos a função `gibbs`. Os parâmetros dessa função são `dados`, os dados ajustados pela função `bnr`, `N`, o número de amostras a serem geradas em cada partição, e `warn`, o número

de iterações necessárias para gerar um aviso sobre o andamento das amostragens.

```
set.seed(12345)
modelo <- d_cjpg %>%
  bnr::bnr(edges, kfolds = 8L) %>%
  bnr::gibbs(1000L, warn = 500L)
```

```
## kfold 01 -------------------------------------
## numero de parametros:  1025
## iteracao:  500
## iteracao:  1000
## kfold 02 -------------------------------------
## numero de parametros:  1025
## iteracao:  500
## iteracao:  1000
## kfold 03 -------------------------------------
## numero de parametros:  1025
## iteracao:  500
## iteracao:  1000
## ...
```

O resultado do modelo é uma lista com k elementos, cada um contendo a iteração entre 1 e k, a base de treino, a base de teste e uma matriz com as amostras para cada parâmetro.

B.2.3 Próximos passos

- Realização de queries.
- Possibilitar inclusão de prioris de forma simples.
- Definir padrões para a priori.
- Trabalhar com bases de dados sem omissão de forma natural.
- Funções para visualização dos resultados.
- Adaptação do pacote `influenced` para tomada de decisão.
- Melhor visualização das redes com o pacote `DiagrammeR`.
- Submeter ao CRAN.

Referências Bibliográficas

Barlow (1987) Richard E Barlow. Using influence diagrams. Relatório técnico, DTIC Document. Citado na pág. 10

Barlow e de Bragança Pereira (1990) Richard E Barlow e Carlos Alberto de Bragança Pereira. Conditional independence and probabilistic influence diagrams. *Lecture Notes-Monograph Series*, páginas 19–33. Citado na pág. 10

Bernoulli (1713) Jakob Bernoulli. *Ars conjectandi*. Impensis Thurnisiorum, fratrum. Citado na pág. 7

Bonassi *et al.* **(2006)** F.V. Bonassi, S. Wechsler, D.K. Colombo e Reginato L.G.M. Relatório de análise estatística sobre o projeto: "análise econômica do direito aplicada a decisões judiciais: o caso dos contratos de arrendamento mercantil para compra de veículos com cláusulas de reajuste associadas ao dólar". Citado na pág. 9

Chen e Ibrahim (2003) Ming-Hui Chen e Joseph G Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, 13(2):461–476. Citado na pág. 46

Dawid (2002) Alexander Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189. Citado na pág. 11

DeGroot *et al.* **(1986)** Morris H DeGroot, Stephen E Fienberg, Joseph B Kadane e Gordon J Apple. *Statistics and the Law*. Wiley New York. Citado na pág. 9

Friedman *et al.* **(2001)** Jerome Friedman, Trevor Hastie e Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin. Citado na pág. 31

Geman e Geman (1984) Stuart Geman e Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741. Citado na pág. 45

Griffin *et al.* **(2010)** Jim E Griffin, Philip J Brown *et al.* Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188. Citado na pág. 47

Hastings (1970) W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109. Citado na pág. 46

Holmes (1897) Oliver Wendell Holmes. The path ofthe law, 10 harv. *L. Rev*, 457:467–68. Citado na pág. 7

Howard (1984) Ronald A Howard. 2005. influence diagrams. ra howard, je matheson, eds. *Readings on the Princinples and Applications of Decision Analysis II*, páginas 719–762. Citado na pág. 3, 10

Hu *et al.* **(2012)** Xiaoxuan Hu, He Luo e Chao Fu. Probability elicitation in influence diagram modeling by using interval probability. *International Journal of Intelligence Science*, 2(04):89. Citado na pág. 11

Lima (2008) Paulo César Ribeiro Lima. Os desafios, os impactos e a gestão da exploração do pré-sal. *Estudo da Consultoria Legislativa da Câmara dos Deputados*. Citado na pág. 14

Little e Rubin (2014) Roderick JA Little e Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons. Citado na pág. 45

Loevinger (1949) Lee Loevinger. Jurimetrics–the next step forward. *Minn. L. Rev.*, 33:455. Citado na pág. 7

Murphy (1999) Kevin P Murphy. A variational approximation for bayesian networks with discrete and continuous latent variables. Em *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, páginas 457–466. Morgan Kaufmann Publishers Inc. Citado na pág. 46

Nunes (2012) Marcelo Guedes Nunes. *Jurimetria aplicada ao direito societário: um estudo estatístico da dissolução de sociedade no Brasil*. Tese de Doutorado, Pontifícia Universidade Católica de São Paulo. Citado na pág. 2

Pearl (1986) Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, 29(3):241–288. Citado na pág. 10

Pearl (2005) Judea Pearl. Influence diagrams-historical and personal perspectives. *Decision Analysis*, 2(4):232–234. Citado na pág. 3

Pearl (2009) Judea Pearl. *Causality*. Cambridge university press. Citado na pág. 9, 11, 30, 41, 42

Ribeiro (1998) Ivan Cesar Ribeiro. Avaliação do risco de ações judiciais: Uma abordagem jurimétrica (risk evaluation of judicial claims: A jurimetric approach). *Available at SSRN 2477006*. Citado na pág. 2

Shachter (1986) Ross D Shachter. Evaluating influence diagrams. *Operations research*, 34(6):871–882. Citado na pág. 10

Spiegelhalter e Lauritzen (1990) David J Spiegelhalter e Steffen L Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605. Citado na pág. 45

Stern e Kadane (2014) Rafael B Stern e Joseph B Kadane. Compensating for the loss of a chance. *arXiv preprint arXiv:1412.1501*. Citado na pág. 9

Von Ahn *et al.* **(2003)** Luis Von Ahn, Manuel Blum, Nicholas J Hopper e John Langford. Captcha: Using hard ai problems for security. Em *Advances in Cryptology—EUROCRYPT 2003*, páginas 294–311. Springer. Citado na pág. 13

Xiaoxuan *et al.* **(2013)** HU Xiaoxuan, JIANG Fan e XIA Wei. Causal influence diagrams for decision-making. *Journal of Convergence Information Technology*, 8(5). Citado na pág. 11

Zabala e Silveira (2014) Filipe Jaeger Zabala e Fabiano Feijó Silveira. Jurimetria: estatística aplicada ao direito= jurimetrics: statistics applied in the law. Citado na pág. 2, 8